Mapping the nuclear microenvironment of genes at a genome-wide scale

Asli Yildirim[1], Nan Hua[1], Lorenzo Boninsegna[1], Guido Polles[1], Ke Gong[1], Shengli Hao[1], Wenyuan Li[2], Xianghong Jasmine Zhou[2], Frank Alber[1♯]

[1]Institute of Quantitative and Computational Biosciences, Department of Microbiology, Immunology, and Molecular Genetics, University of California Los Angeles, 520 Boyer Hall, Los Angeles, CA 90095, USA

[2]Department of Pathology, David Geffen School of Medicine, University of California Los Angeles, 10833 Le Conte Ave, Los Angeles, CA 90095, USA

♯ To whom correspondence should be addressed.

Tel: +1 310-267-0363

Email: falber@g.ucla.edu

Abstract

The folding and subnuclear compartmentalization of chromosomes relative to nuclear bodies is an integral part of gene function. However, mapping the three-dimensional (3D) organization of all genes, in single cells, on a genome-wide scale remains a major challenge. Here, we demonstrate that data-driven population-based modeling—from ensemble Hi-C data alone—can provide a detailed description of the nuclear microenvironment of genes. We define a gene's microenvironment by its subnuclear positions with respect to different nuclear bodies, local chromatin compaction, and preferences in chromatin compartmentalization. These structural descriptors are determined in single cell models on a genome-wide scale, thereby revealing the dynamic variability of a gene's subnuclear microenvironment across a population of cells. We demonstrate that a gene's microenvironment is directly linked to its functional potential in gene transcription, replication and subnuclear compartmentalization. Some chromatin regions are distinguished by their strong preferences to a single microenvironment (either transcriptionally active or silenced), due to strong associations to specific nuclear bodies. Other chromatin shows highly variable microenvironments and lack specific preferences. We demonstrate that our method produces highly predictive genome structures, which accurately reproduce data from TSA-seq, DamID, and DNA-MERFISH imaging. Thus, our method considerably expands the range of Hi-C data analysis.

Introduction

The spatial organization of eukaryotic genomes is linked to regulatory functions of gene transcription, DNA replication, cell differentiation and upon malfunction to cancer and other diseases[1]. Recent advances have led to a prolific development of improved technologies for probing chromosome interactions and 3D organization. Life-cell and super-resolution microscopy[2,3] as well as mapping technologies based on high-throughput sequencing[4-11] shed light into the dynamic formation of chromatin loops, topological associating domains (TADs), and their role in moderating promoter-enhancer interactions between remote DNA regions[12-14]. Besides local and specific chromatin interactions, chromatin function is also controlled by compartmentalization, i.e., preferential associations of chromatin with similar functional profiles. Chromosome conformation mapping shows spatial segregation of chromatin into transcriptionally active and inactive A/B compartments[9], subsequently refined, at high sequencing depth, into 5 primary Hi-C subcompartments[15]. Chromatin compartmentalization is also expressed by associations to nuclear bodies, such as nuclear speckles, PML bodies, Polycomb bodies or lamina associating domains, and other nuclear compartments[16]. Gene positions to nuclear bodies play critical roles in permissiveness of gene expression, as shown by TSA-seq[17] and DamID[6] experiments. Transcriptional permissive regions coalesce at nuclear speckles, nuclear pore complexes and PML bodies, while regions of transcriptional repression are associated with the nuclear lamina and perinucleolar chromatin.

It is increasingly clear that a gene's local microenvironment, defined by its location relative to nuclear bodies and attributes of its adjacent chromatin, influences its functional capacity in gene expression, RNA processing and DNA replication. Several experimental technologies probe the mean distances (TSA-seq) or association frequencies (NAD-seq, DamID) of genes to nuclear speckles[17], lamina associated domains[6], and nucleoli[18]. However, collecting this information simultaneously within the same cell, at the same time, is challenging, especially when considering cell-to-cell variability of a gene's microenvironment within a population of cells. Recent DNA- and RNA-MERFISH super resolution imaging detected, within the same cells, the nuclear locations of 1137 genes, together with the positions of nuclear speckles, nucleoli, as well as the amount mRNA transcripts[19]. However, at this point, the amount of probed genomic DNA regions is sparse, containing ~1% of entire genomes.

Here, we demonstrate the capacity of data driven genome structure modeling to uncover, not only the local microenvironment of genes, but also the dynamic variability of structural features in a population of single cell models. We achieve this goal by using a population-based genome structure modeling approach, which takes *in situ* Hi-C data to generate a population of diploid genome structures statistically

consistent with it[20,21]. We demonstrate that our method produces—from Hi-C data alone—highly predictive genome structures, which reproduce data from SON TSA-seq[17], laminB1 TSA-seq[17], laminB1 pA-DamID[22], GP-seq[23] and 3D FISH experiments. Our models deliver an array of orthogonal descriptors defining a gene's nuclear microenvironment, e.g., its radial position, association frequencies and mean distances to laminB1, speckles and nucleoli, the local chromatin fiber compaction, and local compartmentalization in form of the trans A/B ratio[19]. Moreover, structural descriptors are determined in single cell models, thereby revealing cell-to-cell variability across the population of models. Among all structural descriptors, a gene's speckle association frequency and its trans A/B ratio have the highest predictive value for gene expression. Thus, the relationship between gene expression permissiveness and interior radial position can better be explained by deterministic locations of a gene relative to nuclear bodies, which in turn show stochastic preferences for the nuclear interior. Also, genes with high expression heterogeneity between cells (from single cell RNA-seq data[24]) often show increased cell-to-cell variability in their nuclear microenvironment, indicating that extrinsic noise can contribute to gene expression heterogeneity[25]. Furthermore, Hi-C subcompartments, defined by Rao *et al.*[15], are characterized by unique physical microenvironments, which segregate into a number of spatial partitions, suggestive of processes driven by microphase separation[26].

Although other computational approaches also modeled entire chromosomes or even diploid genomes from Hi-C data[27-33], so far, none documented the predictive accuracy in reproducing multimodal experimental data as presented here. Our findings demonstrate that computational modeling, from Hi-C data alone, produces exceedingly predictive models, providing a detailed description of the subnuclear locations, folding and compartmentalization of chromatin in diploid genomes. Therefore, our approach considerably expands the scope of information retrieved from Hi-C data.

Results

Here, we study 3D structures of diploid lymphoblastoid genomes (GM12878) from *in situ* Hi-C data[15] at 200kb base-pair resolution. Our method generates a population of 10,000 genome structures, in which all accumulated chromatin contacts are statistically consistent with contact probabilities from Hi-C experiments[20,21]. The structure optimization is achieved by solving a maximum likelihood estimation problem utilizing an iterative optimization algorithm with a series of optimization strategies for efficient and scalable model estimation[20,21,34], which accurately reproduce Hi-C contact probabilities (Pearson correlation 0.98, genome-wide; 0.99 and 0.83 for cis and trans contacts, p=~0) (**Extended Data Figs. 1a,c**). More than 99.91% of all contact constraints are fully satisfied and predicted contact frequencies show very small residuals (**Extended Data Fig. 1b**).

Model Assessment. Assessment of model accuracy is achieved by prediction of experimental data not used as input in the modeling process (**Extended Data Table 1**). First, models generated from a sparse Hi-C data set, with 50% entries randomly removed, predict the missing Hi-C contact frequencies with high accuracy (Pearson correlations; cis: 0.93 and trans:0.69 of missing data, p=~0) (**Extended Data Figs. 1d,e**). Second, predicted laminB1 pA-DamID data show good agreement with experiment[22,35] (Pearson's correlation 0.80, p=~0 **Extended Data Figs. 2a,b**), thus reproducing accurately contact frequencies between chromatin and the nuclear envelope. Third, average radial positions of chromatin agree with data from GPseq experiments[23] (Pearson correlation 0.80, p=~0) (**Extended Data Figs. 2a,b**). Fourth, our models predict SON TSA-seq data in excellent agreement with experiment[17], thus predict accurately speckle locations and mean speckle distances (Pearson correlation 0.87, p=~0, **Extended Data Fig. 2a,b**, *Methods*). Fifth, laminB1 TSA-seq data, measuring mean chromatin distances to the nuclear envelope, are predicted with excellent agreement to the experiment[17] (Pearson correlation of 0.78, p=~0, **Extended Data Fig. 2a,b**, *Methods*). Sixth, our models confirm interior radial preferences of chromatin replicated in the earliest G1b phase ($p<1.2e^{-77}$, Mann-Whitney-Wilcoxon one-sided test) and predict a gradual increase in average radial positions for chromatin replicated at progressively later times[36] (**Extended Data Fig. 2c**). Seventh, our models agree with 3D Fluorescent *in situ* Hybridization (FISH) experiments[21,37], namely co-location frequencies of four inter-chromosomal pairs of loci[37] (R = 0.986, F-statistic p=0.01392, **Extended Data Fig. 2d**), and distance distributions between three loci on chromosome 6 and relative differences in radial positions of these loci (**Extended Data Fig. 2e**). Finally, results are reproduced by technical replicates and converge even with smaller population sizes (*Methods, Supplementary Information*).

After establishing the predictive value of our models, we now study folding patterns of chromosomes in the context of nuclear compartments. Specifically, we define a description for a gene's physical microenvironment by its nuclear location, relative distances to nuclear bodies, its interrelation with chromatin compartments, and its variabilities within the cell population (**Fig. 1a**). We then aim to identify chromatin in distinctly specialized microenvironments, determine their spatial segregation and identify the role of microenvironments in regulating transcription and replication. To achieve this goal, we first calculate a variety of structural features for each chromatin region (see **Fig. 1b** for an overview of all structural features discussed in this paper).

Subnuclear gene positions. Radial positions of genes are of functional relevance. DNA FISH experiments revealed for a number of genes, upon transcriptional activation, a statistical shift of their locations towards the nuclear center[38,39] [17]. Within a chromosome, average radial positions (relative to the nuclear center) reveal deep minima (**Fig. 2a**), which overlap with regions of lowest laminB1 scDamID signals **(Extended Data Fig. 2f**.)[35] (93% of chromatin with the 25% lowest average radial positions show either no detectable or only occasional contact (<20% frequency) with lamina[35]) (**Extended Data Fig. 2g**.) These minima are often flanked by regions undergoing large radial transitions over short sequence distances (i.e., regions of high radial gradient) (**Fig. 2a**). Such transitional regions often align remarkably well with Hi-C subcompartment borders[15]: 76% of regions with high radial gradient overlap with subcompartment borders (**Fig. 2a**, *Methods*). Rao *et al.*[15] divided chromatin into 5 primary Hi-C subcompartments—two transcriptional active (A1,A2) and three inactive subcompartments (B1,B2,B3). A1 chromatin, gene dense with relatively high GC content, show the lowest—most interior—average radial positions with a narrow distribution (**Fig. 2c**). The highest probability is observed at the most interior radial shell, with sharply decreasing probabilities otherwise (**Fig. 2b**). A2 chromatin display not only larger, more exterior, average radial positions but overall do not share a common preference—thus has a more wide-spread distribution with more even probabilities across all nuclear shells (**Figs. 2b,c**). A similar behavior, with a relatively wide distribution is also seen for B2, enriched in pericentromeric and nucleoli associated domains (NADs), with a slight increase in probabilities towards the outer half of nuclear shells (**Fig. 2b**). The B1 subcompartment, although transcriptionally silenced (Polycomb-related heterochromatin), shares similar location preferences to A1—with highest probabilities in the interior radial shells, gradually decreasing outwards (**Figs. 2b**). Radial positions alone are evidently not sufficient indicators of transcriptional activity. B3 chromatin is most peripheral, with a relatively narrow distribution in average positions (**Fig. 2c**). 80% of B3 chromatin are lamina-associated domains (LADs)[15], mostly heterochromatin, confirming its probabilities at the outermost two shells with decreasing values towards the interior (**Fig. 2b**). Our results confirm similar positional preferences from GPseq data analysis[23].

Cell-to-cell variability of gene positions. Due to their stochastic nature, radial positions of the same locus vary between cells. Some loci are in an interior position in one structure and close to the periphery in another (**Fig. 2d, lower panel)**. To illustrate these stochastic fluctuations, we calculated δ_{RAD}, the log-transformed fraction of observed and expected standard deviations of a gene's radial positions ($\delta_i^{RAD} = log_2(\sigma_i/\langle\sigma\rangle)$ (*Methods*). δ_{RAD} differs distinctly between loci and is a decisive indicator to distinguish subcompartments (**Fig. 2d**). Continuous blocks of chromatin show similar trends in δ_{RAD} and are often part of the same subcompartment. Blocks of high variability ($\delta_{RAD} > 0$) alternate, in sharp transition, with blocks of low variability ($\delta_{RAD} < 0$)—transitions between high and low variability regions often coincide with subcompartment borders, most prominently between A2 and B3 subcompartments (**Fig. 2d**). A1 and A2 subcompartments, both active, can be distinguished by their structural variability alone (**Fig. 2d**): A1 loci show overall the lowest, and A2 one of the highest δ_{RAD} values (**Fig. 2e**.) 93% of all highly variable regions ($\delta_{RAD} > 0$) in the active compartment are A2 chromatin. Interestingly, A2 regions with the largest variability often show a bimodal distribution in radial positions, i.e. the same gene is located, in different models, in two states—in a fraction of models the gene is in the nuclear interior, while in the remainder it resides towards the periphery (**Fig. 2f,g**). We hypothesize that such genes may exist, in the population, in two functional states: active in the transcriptional favorable interior, and otherwise in the predominantly silencing environment at the periphery. Indeed, the most highly variable A2 regions (HV) are significantly more enriched in repressive H3K9me3 histone modifications than other A2 regions, while maintaining some enrichment for active histone modifications (H3K27ac, H3K9ac), but to a lesser extent than low variable (LV) A2 regions. However, highly variable A2 regions are significantly depleted in H3K27me3 in comparison to all other A2 regions (**Fig. 2h**). Additionally, transcribed genes with the lowest number of transcripts in single cell RNA-seq (scRNA-seq)[24], show significantly higher variability in radial positions (δ_{RAD}) and are in the transcriptionally favorable central regions in a smaller percentage of models than the group of genes with the highest number of transcripts (**Fig. 2i**).

Spatial fragmentation of subcompartments. Chromosome folding permits chromatin in the same subcompartment, separated in sequence or by chromosomes, to congregate into a number of spatial partitions (**Fig. 3a**). To analyze subcompartment fragmentations, we represent each structure as a chromatin interaction network (CIN), with chromatin regions as nodes connected by edges if the chromatin is in physical contact (**Fig. 3b**, *Methods*). A CIN is constructed separately for each subcompartment and each structure. CINs show a heterogeneous organization, (i.e. contain a number of highly connected subgraphs), thus are divided spatially into a number of local partitions with enhanced neighborhood connectivity—an organization reminiscent to microphase fragmentations, instigated by the

physical nature of the chromatin polymer preventing the segregation of each subcompartment into a single macrophase.

To identify highly connected subgraphs we apply the Markov Clustering Method (MCL) to each CIN[40]. The locations of dense subgraphs in a 3D structure define subcompartment partitions, areas with highest local concentration of chromatin in a subcompartment, visualized by their occupied volume in the genome structures (**Fig. 3b,c**). The size, number and locations of spatial partitions vary between subcompartments (**Figs. 3c,d,e Extended Data Table 2**). A1 networks have the highest neighborhood node connectivity and number of maximal cliques (p=~0, Mann-Whitney-Wilcoxon one-sided test, **Extended Data Table 2**) and subsequently are fragmented into the lowest number of partitions (54±5 per structure) with the largest average sizes (71±7 nodes (~14Mb)) (**Fig. 3d,e, Extended Data Table 2**). They contain the highest fraction of inter-chromosomal interactions (42%) (**Fig. 3f**). A2 networks show higher fragmentations into larger numbers of smaller partitions (33±2 nodes), which are dominantly formed by intra-chromosomal interactions (75%) (**Extended Data Table 2, Fig. 3d,e,f**). Like A2, B1 networks also show high fragmentations (92±6 per structure) into relatively small partitions (33±2 nodes) (**Fig. 3d,e**). However, B1 partitions are formed by a large fraction of inter-chromosomal interactions (35%) (**Fig. 3f**). B3 partitions are relatively large and formed by the highest fraction of intra-chromosomal interactions (90%) (**Fig. 3e,f**).

There are two important criteria to examine: First, are partitions homogenous in their subcompartment composition, and second, are partitions of different subcompartments packed in a preferential order in the nucleus? To investigate these questions, we calculate the neighborhood composition enrichment, as the log ratio of observed/expected fractions of subcompartment chromatin located in the immediate neighborhood of chromatin regions (**Fig. 3g**, *Methods*). A1, B2 and B3 partitions are most homogenous—chromatin is preferentially surrounded by their own kind (strong diagonal in neighborhood composition matrix, **Fig. 3g**). B1 and A2 chromatin show high neighborhood enrichment to A1 and B3, respectively, partly due to their higher fragmentation into smaller partitions. B1 partitions associate with, and even surround, A1 partitions, confirming their similar radial profiles and enhanced SON TSA-seq signals for silenced B1 chromatin[17,23], while A2 chromatin is also observed in B3 neighborhoods (**Fig. 3g,h**).

To assess the functional relevance of partitions, we investigate gene expression levels with respect to their locations in the partition. We averaged the nascent RNA expression (from GRO-Seq experiments[41]) for all genes located in concentric shells around partition centers. A1 partitions show the highest transcriptional activities towards the partition centers, with decreasing activities to outer regions (**Fig. 4a**). Thus, A1 partitions are regional hubs with highest transcriptional activities. A2 partitions show similar trends, although substantially lower signals (**Fig. 4a**). Expression levels at centers of large A1 and A2

partitions are significantly higher than those of smaller partitions and most highly expressed genes reside preferably in larger partitions (**Fig. 4b**).

Predicting nuclear speckle locations. Centers of A1 partitions share several attributes with nuclear speckles[42-48] in terms of their numbers per structure, interior positions[49] and associations with chromatin of highest transcriptional activity[50]. Also, mean cytological distances to speckles, from SON TSA-seq experiments, are generally smallest for A1 chromatin[17].

Indeed, averaged TSA-seq signals are strongest at central regions of A1 partitions, and are decreasing towards outer regions (**Fig. 4c**), suggesting that A1 partition centers can serve as approximate point locations for speckles in our models. Despite transcriptional activity, A2 partitions are devoid of TSA-seq signals, while central areas of all other subcompartment partitions are depleted in signals (**Fig. 4c**). To assess if A1 partition centers represent individual speckles, we simulate the experimental TSA-seq process in our models (**Fig. 4d**). SON-TSA produces a gradient of diffusible tyramide free-radicals, instigated at speckle locations, for distance-dependent biotin labelling of DNA[17]. The diffusion and steady state concentration of tyramide free-radicals, at any given nuclear position, is then modeled with an exponential decay function from the distances to all predicted speckles[17] (**Fig. 4d**, *Methods*). The simulated SON TSA-seq signal is then calculated for each chromatin region, in each cell model and averaged over all models in the population, mimicking the TSA-seq experimental procedure in a population of cells. The genome-wide TSA-seq data, predicted from our models, agrees remarkably well with experiment (Pearson 0.87 p=~0) (**Fig. 4e**, **Extended Data Fig. 2a**), capturing peak sizes and distributions. For instance, the TSA-seq profile of chromosome 2 is reproduced with high correlation (Pearson 0.90, p=~0), even though it contains only few A1 regions (6.4%) (**Fig 4e**). We also calculated the enrichment of epigenetic and functional features for chromatin divided into deciles of predicted TSA-seq signals. The agreement with the corresponding analysis from experimental TSA-seq data[17] confirms high prediction accuracy across all ranges of TSA-seq values (**Extended Data Fig. 3a**). Moreover, our models confirm the proposed correlation between mean speckle distances and experimental TSA-seq data (**Extended Data Fig. 3b**).

Speckle locations can be accurately predicted even without A1 annotations, by determining spatial partitions from chromatin with the 10% lowest average radial positions (78% of chromatin with 10% lowest radial positions are part of A1.) On average, these partition centers are within 500nm to those derived from A1 partitions in 99% of structures, and predicted SON TSA-seq data is almost identical with excellent accuracy (Pearson correlation 0.86, p=~0) (**Extended Data Fig. 3c, Extended Data Table 1**).

TSA-seq predictions require correctly folded genomes—the prediction accuracy decreases when TSA-seq data is simulated from A1 sequence positions, or random chromosome conformations (Pearson

0.35, and 0.60, respectively, p=~0, **Extended Data Fig. 3d, Extended Data Table 1)** Similarly, SON TSA-seq data simulated from A2 partitions fail to predict TSA-seq data (Pearson 0.18, $p=9.4x10^{-98}$ **Extended Data Fig. 3c**), although A2 partitions are transcriptional hubs, suggesting different principles of formation than speckle associated chromatin partitions.

Interestingly, TSA-seq signals predicted from isolated chromosomes, in identical conformations but deprived of their nuclear context, show dramatically reduced accuracy for some chromosomes. For chromosome 17, the Pearson correlation between experiment and prediction drops from 0.82 to 0.52 without trans-contributions. Also genome-wide predictions accuracy was reduced (Pearson 0.73, p=~0, **Extended Data Fig. 3d, Extended Data Table 1**).

The predicted TSA-seq profiles show a relatively high correlation (Pearson 0.92, p=~0, **Fig. 4f**) with the interchromosomal contact probability (ICP), calculated from our models, as the fraction of its trans vs. all interactions, confirming speckles as major hubs of inter-chromosomal interactions of active regions[17,51].

Defining speckle associated features. Using predicted speckle locations as reference points we can now define additional structural features of a chromatin region's microenvironment, namely (i) its mean distance to the closest speckle, (ii) the cell-to-cell variability of its speckle distance, and (iii) its speckle association frequency (SAF), as the fraction of models a chromatin region is in close association with a speckle (**Fig. 1**, *Methods*).

In a recent landmark study, Su *et al.*[19] used MERFISH microscopy to measure SAF and speckle distances for 1,137 genes in ~5,000 IMR-90 cells. The SAF predicted from our models agrees remarkably well with those from this study (Pearson 0.79, $p=8.4x10^{-223}$, **Fig. 4g**, *Methods*). Moreover, Su *et al.* demonstrated that a gene's SAF is correlated with its trans A/B density ratio, defined as the ratio of A and B compartment chromatin forming inter-chromosomal interactions with the target loci[19]. Trans A/B ratios calculated from our models also show good agreement with experiment (Pearson 0.70, $p=7.6x10^{-109}$, **Fig. 4h**), confirming in GM12878 cells a very high correlation between trans A/B ratios and SAF (Pearson 0.98, p=~0 **Fig. 4i**)[19].

Our models also allow a structural interpretation of previously described TSA-seq trajectories, steep transition in the TSA-seq profile from low to high peaks and back, over relatively short sequence distances (**Fig. 4j,** top panel)[17]. In our models TSA-trajectories correspond to steep transitions in a chromosome's average radial position profile (**Fig 4j**, lower panel). In individual models, these chromosome regions loop from the outer nuclear zone towards the nuclear interior to associate with a nuclear speckle at the loops apex before looping back out (**Fig. 4k**), confirming similar observations in FISH experiments by the Belmont laboratory[17].

SON TSA-seq experiments identified two types of transcription "hot zones": Type 1 regions with high and type 2 with intermediate SON TSA-seq signal peaks[17]. Our models confirm the expectation that type 1 regions have significantly smaller mean speckle distances than type 2 (Mann-Whitney-Wilcoxon two-sided test, $p=1.3x10^{-51}$, **Fig 4l**). However, TSA-seq data is inconclusive whether type 2 regions persistently reside at intermediate speckle distances or localize at speckles in a small fraction of cells and far from them in others[17,50]. Our models uncover the latter case. The vast majority of Type 2 regions show a significantly higher variability in radial positions (**Fig. 4m**) and speckle distances (Mann-Whitney-Wilcoxon two-sided test, $p=1.94x10^{-43}$), associate to speckles in only a smaller fraction of cells (average SAF < ~17%) and thus do not reside stably at intermediate speckle distances (**Fig. 4l**). Type 2 regions show a wide and, in many cases, bimodal distribution of mean speckle distances (**Fig. 4n**). In contrast, Type 1 regions show rather stable radial positions at close speckle distances and a single peak in the mean speckle distance distribution (**Fig. 4n**) resulting in high SAF of at least 50% on average (Mann-Whitney-Wilcoxon two-sided test, $p=2.8x10^{-53}$, **Fig. 4l**).

Transcriptional activity is correlated to speckle association. We now investigate in more detail if a gene's local microenvironment echoes its functional activity. We compare the stochastic variability of gene-speckle distances in single cell models, plotted by the heatmap of each gene's cumulative distance distributions (**Fig. 5a** top panel), with the variability of single cell gene expression, plotted by the heatmap of each gene's cumulative distribution of mRNA transcript counts from scRNA-seq data[24] (**Fig. 5b,** top panel). The two distributions show striking similarities. Moreover, the scRNA-seq transcription frequency (TRF) (the fraction of cells a transcript is detected[24]) and SAF profiles are remarkably similar (**Fig. 5a,b**, lower panels) and show highly significant correlation (**Fig. 5c**, Spearman 0.51, p=~0). Genes with transcripts in a large fraction of cells are located close to speckles in a large fraction of models. We also confirmed our observations with RNA-MERFISH data, which measured for 1,137 genes the fraction of cells where nascent RNA transcripts of each gene is detected (TRF)[19] (**Fig 5d**). Also here, we observe the identical highly significant correlation between TRF and SAF (Spearman 0.51, $p=1.6x10^{-64}$). Interestingly, the correlation between TRF and SAF is substantially larger than the correlation of TRF with a gene's interior location frequency (ILF), for both scRNA-seq and RNA-MERFISH data (Spearman 0.42, p=~0 (scRNAseq) and 0.45, $p=4.1x10^{-50}$ (RNA-MERFISH)) (**Fig 5d**), a possible indication that the preferred interior positions of activated genes may be a consequence of favored associations with nuclear speckles, which themselves show stochastic preferences towards the nuclear interior[17,50,52].

Next, we explore which structural feature is most discriminative in separating actively transcribed genes with the 10% highest from those with the 10% lowest transcript counts. We assess the following features of each gene: the (i) average speckle distance, (ii) cell-to-cell variability of average speckle distance, (iii)

ILF, (iv) SAF, (v) average radial position, (vi) cell-to-cell variability of average radial position, and (vii) median trans A/B ratio[19]. The distributions of all feature values show substantial differences between the two groups of genes (**Fig. 5e**). However, SAF and the highly correlated trans A/B ratios, again, outperform all other features in distinguishing highly from lowly expressed genes, as shown by the highest area under the receiver operating characteristics (ROC) curve (**Fig. 5f**).

Defining nucleoli and lamina associated features. To complete the description of a gene's nuclear microenvironment, we also calculate structural features in relation to nucleoli, the repressive lamina compartment as well as local structural properties of the chromatin fiber.

LaminB1 associated features are calculated using the nuclear envelope as reference point (*Methods*). LaminB1 TSA-seq and DamID data, simulated from our models are in good agreement with experiment[17] (**Extended Data Fig. 2a, Extended Data Table 1**), thus our models describe accurately the mean distances and contact frequencies of genes with the lamin compartment. Lamina association frequencies (LAF), in our models also show good agreement with those from DNA-MERFISH[19] (Pearson 0.64, $p=\sim 3.6\times10^{-119}$) (**Fig. 5g**), although the correlation is lower than for SAF predictions, likely due to the shape differences between flat ellipsoid IMR-90 and spherical GM12878 cell nuclei. Predicted LAF values are inversely correlated with a gene's trans A/B ratios and SAF, confirming previous observations[19] (**Fig. 5h**).

To calculate features in relation to nucleoli, we determine spatial partition centers of chromatin known to be nucleolus organizing regions (NOR) (short arms of chromosomes 13,14,15,21 and 22), and nucleolus associated domains (NADs)[53] (*Methods*), which then serve as reference points to calculate a gene's mean nucleoli distance, its cell-to-cell variability, nucleoli association frequencies (NAF) and nucleoli-TSA-seq data (**Fig. 1**). The NAF calculated in our models shows good agreement with NAF extracted from MERFISH imaging (Pearson 0.71, $p=1.2\times10^{-152}$, **Fig. 5i**, *Methods*)

Finally, we also calculate features of the chromatin fiber. The volume occupied by a chromatin region relates to its local compaction and is estimated, for each chromatin region, by the radius of gyration (RG) of a continuous 1Mb window centered at the target locus (**Extended Data Fig. 4a**, *Methods*). Average RG profiles show pronounced maxima at locations of TAD boundaries, while minima show domain-like compaction (**Extended Data Fig. 4a,c,d**). RG profiles vary between cells (**Extended Data Fig. 4b**) and the probability for observing a peak is at maximum at TAD border locations, while randomly selected regions show a flat probability distribution (**Extended Data Fig. 4e)**. About 20% of structures show a RG peak (i.e., domain border) at the exact TAD border location (50% show a RG peak within the immediate vicinity). These TAD border frequencies in single cell structures agree with recent oligoSTORM superresolution imaging[2].

The spatial microenvironment of a gene mirrors its functional state. Overall, we characterized the nuclear microenvironment of each genomic region by a total of 17 structural features (**Fig. 1**), covering global features of nuclear organization, local properties of the chromatin fiber and the dynamic variability of these features between models. We now assess if the nuclear microenvironment can explain functional differences between chromatin.

Chromatin in Hi-C subcompartments are distinct in their enrichment patterns of structural features, thus they represent well defined physical microenvironments (*Methods*) (**Fig. 6a,b, Extended Data Fig. 5**). The most discriminating feature is the SAF. A1 and B3 show strongly anti-correlated enrichment patterns across all features, except δ_{RAD}—both show low variability in their nuclear positions. Generally, A1 and B3 chromatin show high uniformity within their class with small dynamic variations between models. A2 regions are very different, with relatively weak enrichement patterns and unusually high cell-to-cell variability in radial locations, speckle distances and overall wide distributions of feature values within their class, indicating no clear location preferences with respect to nuclear bodies (**Fig. 6a, Extended Data Fig. 5**). Also, some A2 regions show multi-state properties—they share microenvironments reminiscent of one subcompartment in some models, while features of another state in others. B2 and B3 chromatin are clearly separated by their microenvironment, mostly based on nucleolar and lamina associated features (**Fig. 6a,b**). However, B2 chromatin is distinct in its high variability of nuclear locations, possibly explained by prevalent locations of nucleoli at both central and peripheral regions (**Fig. 6a,b)**. B1 chromatin, linked to polycomb bodies, are quite different from any other inactive subcompartment (**Fig. 6a**). Instead, B1 shares similar microenvironment with highly active A1, although with substantial smaller enrichments. Thus, silenced B1 genes would be in a position of highest transcriptional potency, if activated.

To demonstrate that a gene's microenvironment embodies functional information, we predicted Hi-C subcompartments from structural features alone: K-means clustering of active chromatin based on their structural microenvironment predicts A1 and A2 subcompartments with 94% accuracy, while chromatin in inactive subcompartments were predicted with an accuracy of 84%, comparable in accuracy to supervised methods using Hi-C contact frequencies[54] (*Methods*, **Fig. 6c**).

A gene's microenvironment also reflects its transcriptional potential. Genes with the top 10% highest expression levels (T10) are clearly distinguished in their microenvironment from genes with the bottom 10% expression levels (B10) (**Fig. 6d,e**). T10 genes show very strong enrichment patterns, thus, are preferentially located in specific locations in the nucleus, in particular in relation to nuclear speckles (**Fig. 6e**). Lowly expressed B10 genes do not show any preferred localization patterns with no preferential positioning relative to nuclear bodies, and more variable nuclear locations (**Fig. 6e**).

Enhancers (EN) and superenhancers (SEN) show similar trends in enrichment patterns as T10 genes (**Fig. 6f**). However, SEN have substantially higher enrichment and depletions in structural features, indicating a stronger preference in their microenvironment, in particular for higher SAF, interior positions, transA/B ratio, ICP and depletion of LAF values (**Fig. 6f**). Notably, for both EN and SEN features related to cell-to-cell dynamic variability are depleted in comparison to the genome-wide average.

The structural microenvironment of genomic regions is also linked to replication timing. A gene's microenvironment changes gradually with increasing replication timing (**Fig. 6g.**). Chromatin replicated in early phases (G1b, S1) show similar patterns to highly active genes, enriched in SAF, interior regions, with relatively low structural variability. The S2 phase comprises chromatin located towards the interior without enrichment at nuclear speckles, while chromatin in S3, S4 and G2 are depleted in the interior and enriched in lamina associated features. Notably, chromatin replicating in the S2 and S3 phase shows the highest dynamic variability in their nuclear positions and are not preferentially associated to nuclear bodies. Overall, the most discriminative feature of different phases is SAF and transA/B ratio.

Finally, chromatin divided by TSA-seq values into 10 groups shows distinguished microenvironment and gradually changes enrichment patterns with increasing TSA-seq values (**Fig.6i,j**). Chromatin in the first (d1,d2) and last (d9, d10) deciles show the most stable microenvironment with depleted variability (**Fig. 6j**). In contrast, chromatin in deciles d4-d7 are structurally less defined, highly variable in nuclear positions and decile 6 shows no preferred locations towards nuclear bodies.

Discussion

Here we introduce data driven genome structure modeling to map the nuclear microenvironment of genes on a genome-wide scale. Our approach is unique in that it determines all structural features simultaneously for each gene in single cell models, thus allowing to capture the dynamic variability of a gene's microenvironment in a population of models.

The nuclear microenvironment of a gene can be a good indicator of its replication timing, subcompartment associations and transcriptional potential. For instance, the frequency of close speckle associations appears as an important factor in a gene's transcriptional potency[55]. Chromatin with the 10% highest and lowest transcriptional activity can be distinguished based on their distinct feature enrichments. Chromatin replicated at the earliest phase are distinct in their microenvironment from those replicating at latest stages.

There are several other interesting observations. Speckles appear to be the single hub of inter-chromosomal interactions of active chromatin regions, confirming observations from SPRITE experiments[51]. Moreover, the preferred interior positions of activated genes could be a consequence of preferential positions relative to nuclear speckles, which in turn have a stochastic preference towards the nuclear interior[17]. Our observations also confirm that Hi-C subcompartments define physically distinct chromatin environments.

Overall, we observe two major categories of chromatin. For one, chromatin that is strongly associated to a single microenvironment, either transcriptionally active or silenced—they are well defined by their strong associations to specific nuclear bodies in the majority of cells with only little cell-to-cell variability (even though absolute locations in the nucleus may vary). Among those are chromatin of the A1 and B3 subcompartment. These regions show a strong preference in nuclear locations and subsequently strong enrichment and depletion patterns in their structural features. These regions are most homogenous in functional properties within their respective state, leading to the highest expression rates, earliest and latest replication. Regions containing superenhancers are also part of this group.

The second type of chromatin is characterized by the lack of particular preferences in their locations relative to nuclear bodies. These genes appear highly variable in their nuclear positions between cells, have intermediate replication timing (phases S2, S3) and if actively transcribed show relatively low transcript frequencies, low interchromosomal contact probability and trans A/B ratios in comparison to actively transcribed gene in the first category. In TSA-seq experiments most of such regions were identified as type II peak regions, with intermediate TSA-seq values. We also noticed that active regions in this category form relatively small spatial subcompartment partitions (i.e., microphases) dominated by

intra-chromosomal interactions, in contrast to the larger spatial partitions, dominated by inter-chromosomal interactions, observed for speckle associated active chromatin. Among those regions are chromatin of the A2 subcompartment.

Data-driven genome modeling can provide rich information not directly accessible from the Hi-C data. Our models, from Hi-C data alone, predict with good accuracy the SAF and trans A/B ratio of chromatin from superresolution imaging, as well as data about mean distances to nuclear bodies from TSA-seq experiments. Our method considerably expands the range of Hi-C data analysis. This is important as Hi-C data is readily available for a multitude of cells and tissues and a comparative analysis of a gene's microenvironment can be a powerful tool for structure function studies.

Our method also has its limitations, which we will address in future work. Currently, the nuclear bodies are represented without excluded volumes. However, in its current form we can demonstrate by the accuracy of our predicted features, that our methods produce a first approximation for a multitude of structural features that can be highly relevant for a better understanding of genome structure function relationships. In future, we plan to incorporate nuclear shapes from imaging into the modeling process.

Methods

Population Based 3D Structural Modeling

- **General description**

Our goal is to generate a population of 10,000 diploid genome structures, so that the accumulated chromatin contacts across the entire population are statistically consistent with the contact probability matrix $\mathbf{A} = (A_{ij})_{N\times N}$ derived from Hi-C experiments[21,37]. To achieve this goal, we utilize population-based modeling, our previously described probabilistic framework to de-multiplex the ensemble Hi-C data into a large population of individual genome structures of diploid genomes statistically consistent with all contact frequencies in the ensemble Hi-C data[20,21].

The structure optimization is formulated as a maximum likelihood estimation problem solved by an iterative optimization algorithm with a series of optimization strategies for efficient and scalable model estimation[20,21,34]. Briefly, given a contact probability matrix $\mathbf{A} = (A_{ij})_{N\times N}$, we aim to reconstruct all 3D structures $\mathbf{X} = \{\boldsymbol{X_1}, \boldsymbol{X_2} \ldots \boldsymbol{X_M}\}$ in the population of M models, each containing 2*N* genomic regions for the diploid genome (at 200 kb base-pair resolution), and $X_{im} \in \Re^3, i = 1..2N$ as coordinates of all diploid genomic regions in model M. We introduce a latent indicator variable $\mathbf{W} = (\boldsymbol{w_{ijm}})_{\mathbf{2N\times 2N\times M}}$ for complementing missing information (i.e. missing phasing and ambiguity due to genome diploidy). $\mathbf{W}$ is a binary-valued 3rd-order tensor specifying the contacts of homologous genomic regions in each individual structure of the population, such that $\sum_{m=1}^{M} \boldsymbol{W}_m/M = \boldsymbol{A}$. We can jointly approximate the structure population $\mathbf{X}$ and the contact tensor $\mathbf{W}$ by maximizing the log-likelihood of the probability:

$$\log P(\mathbf{X}|\mathbf{A}, \mathbf{W}) = \log P(\mathbf{A}, \mathbf{W}|\mathbf{X})$$

$$subject\ to \begin{cases} nuclear\ volume\ confinement \\ excluded\ volume \\ chain\ connectivity\ restraint \end{cases}$$

where

i. Nuclear volume constraint: All chromatin spheres are constrained to the nuclear volume with radius R_{nuc}; $\|\vec{x}_{im}\|_2 \leq R_{nuc}$, where $\|\vec{x}_{im}\|_2$ is the distance of the region i from the nuclear center in structure m.

ii. Excluded volume constraint: This constraint prevents overlap between two regions represented by spheres, defined by their excluded volume radii (R^{ex}); $\left\|\vec{x}_{im} - \vec{x}_{jm}\right\|_2 \geq 2 \times R^{ex}$.

iii. Polymer chain constraint: Distances between two consecutive 200-kb spheres within the same chromosomes are constrained to their contact distance to ensure chromosomal chain integrity; $\left\|\vec{x}_{(i+1)m} - \vec{x}_{im}\right\|_2 \leq 2 \times R^{soft}$, where $R^{soft} = 2 \times R^{ex}$.

Our modeling pipeline uses a step-wise iterative process, in which the optimization hardness is gradually increased by adding contacts with decreasing contact probabilities in the input matrix. The iterative optimization procedure involves two steps, each optimizing local approximations of the likelihood function: (1) Assignment step (A-step): Given the estimated structures $\mathbf{X}$ at step k, estimate $\mathbf{W}$; and (2) Modeling step (M-step): Given the estimated $\mathbf{W}$, generate model population $\mathbf{X}$ at step k+1 that maximizes likelihood to observe $\mathbf{W}$. Structures in the M-step are calculated using a combination of optimization approaches, including simulated annealing molecular dynamics simulations.

Moreover, during each optimization cycle we also use iterative refinement steps, a methodological innovation for effective reassignment of restraints during the optimization process, which allows genome structure generation at higher resolution and improved accuracy in comparison to our previous approach[20,21] (see Iterative refinement method in *Supplementary Information*).

After 11 iterations, our method converges and the genome-wide contact probabilities from the structure population agree remarkably well with those from the Hi-C experiment.

▪ Genome representation

The nucleus is modeled as a sphere with 5 μm radius (R_{nuc})[21]. Chromosomes are represented by a chromatin chain model at 200-kb base-pair resolution. Each 200-kb chromatin region, in the diploid genome, is modeled as a sphere, defined by an excluded volume radius (R^{ex} = 118 nm). R^{ex} is estimated from the sequence length, the nuclear volume and the genome occupancy (40%), as described in ref.[21]. The full diploid genome is represented with a total of 30,332 spheres.

<u>Random starting configurations.</u>

Optimizations are initiated with random chromosome configurations. Chromatin regions are randomly placed in a bounding sphere proportional to its chromosome territory size and randomly placed within the nucleus.

- **Comparison between contact frequency maps from Hi-C experiment and model population.**

To quantify the agreement between Hi-C experiment and model population, we perform the following analyses:

1) Comparison between input and output Hi-C maps are evaluated by Pearson correlation coefficients.

2) Restraint violation ratios. On average about 175,304 contact restraints are imposed in each of the 10,000 structures. The restraint score of each contact restraint i is calculated as: $Violation\ ratio_i = \frac{d_i - D}{D}$, where d_i is the distance between the contact loci in the model, and D is the target contact distance ($2 \times R^{soft}$).

3) Residual ratio. The residual ratio Δr is defined as:

$$\Delta r_{kl} = \left(f_{kl}^{input} - f_{kl}^{model}\right) / f_{kl}^{input}$$

with f_{kl}^{input} and f_{kl}^{model} as the contact probabilities between regions *k* and *l* from experiment and models, respectively. Residual ratios are very small, and centered at a median of 0.03 (mean= -0.05) for intra-chromosomal and 0.001 (mean = -0.002) for inter-chromosomal contacts (Fig. S1), showing excellent agreement between experiment and model.

4) Prediction of missing Hi-C data from sparse data model. A sparse Hi-C input data set is generated by randomly removing 50% of the non-zero data entries from the Hi-C contact frequency matrix.

- **Robustness and Converge Analysis**

<u>Replicates</u>

Technical replicates are calculated from different random starting configurations. Resulting contact frequency maps and the average radial positions of all chromatin regions between replica populations are nearly identical (Fig. S2). All observed structural features discussed in this paper are reproduced in the technical replicate population.

<u>Population size</u>

To test convergence with respect to population size, we generate 5 different populations with 50, 100, 1,000, 5,000 and 10,000 structures. Chromatin contact frequencies and structural features for each structure populations are compared against results with a population size of 10,000 structures. At a population of 1,000 structures, a size much smaller than our target population, contact frequency values

and average radial positions are already converged to a very high correlation with those from a 10,000 structure population (Fig. S3).

Chromatin interaction networks and identification of spatial partitions

- **Building chromatin interaction networks**

A chromatin interaction network (CIN) is calculated for each model and for chromatin in each subcompartment separately as follows: Each vertex represents a 200-kb chromatin region. An edge between two vertices *i, j* is drawn if the corresponding chromatin regions are in physical contact in the model, if the spatial distance $d_{ij} \leq 2 \times (R^{soft})$.

<u>Network properties</u>

Maximal Clique Enrichment: A clique is a subset of nodes in a network where all nodes are adjacent to each other and fully connected. The maximal clique refers to the clique that cannot be further enlarged. The number of maximal cliques, c, is calculated using the *graph_number_of_cliques* function in the *NetworkX* python package[56]. The maximal clique enrichment (MCE) of the subcompartment *s* in the structure *m* is calculated as:

$$MCE_{s,m} = \frac{c_{s,m}}{\frac{1}{10}\sum_{r=1}^{10} c_{r,m}}$$

Where $c_{s,m}$ is number of maximal cliques for subcompartment s in structure m; $c_{r,m}$ is the number of maximal cliques of a CIN constructed from randomly shuffled subcompartment regions in the same structure m. High MCE values shows formation of a structural subcompartment with high connectivity between 200-kb regions of the same state.

Neighborhood Connectivity: To calculate the neighborhood connectivity (NC) of a subcompartment CIN, we first calculate the average neighbor degree for each node using the *average_neighbor_degree* function in the *NetworkX* python package[56]. The overall neighborhood connectivity of the subcompartment s in the structure m is then calculated as:

$$NC_{s,m} = \frac{1}{N_{s,m}} \sum_{j=1}^{N_{s,m}} deg_j$$

where $N_{s,m}$ is the number of nodes in the CIN of the subcompartments s in the structure m , and deg_j is the average neighbor degree of node *j*.

▪ Identifying spatial partitions via Markov clustering

Spatial partitions of subcompartments are identified by applying Markov Clustering Algorithm (MCL)[40], a graph clustering algorithm, which identifies highly connected subgraphs within a network. MCL clustering is performed for each subcompartment CIN in each structure by using the *mcl* tool in the *MCL-edge* software[40]. Unless otherwise noted, the 25% smallest subgraphs (with less than 7 nodes, many of those singletons) are discarded from further analysis to focus on highly connected subgraphs. The highly connected subgraphs are referred to as "spatial partitions" throughout the text.

In addition to subcompartment partitions, we also predict speckle, and nucleoli partitions as follows:

i. Speckle partitions:

Case 1: *Predictions of speckle locations with knowledge of A1 subcompartment annotations*

Speckle locations are identified as the geometric center of A1 spatial partitions identified by Markov clustering of A1 CINs. In each structure, A1 spatial partitions are considered with sizes larger than 3 nodes (chromatin regions).

Case 2: *Predictions of speckle locations without knowledge of subcompartments*

We first identify chromatin expected to have high speckle association. These regions are identified as those with unusually low and stable interior radial positions. We select 10% chromatin regions with the lowest average radial positions. (78.4% of these regions are part of the A1 subcompartment). We then generate CINs for the selected group of chromatin regions in each structure of the population. Approximate speckle locations are then identified as the geometric center of the resulting spatial partitions identified by Markov clustering of the CINs. Spatial partitions are considered with sizes larger than 3 chromatin regions.

Case 3: *Predictions using locations of A2 partition centers*

For comparison, we also identify speckle locations as the geometric center of A2 spatial partitions identified by Markov clustering of A2 CINs similar to Case 1. In each structure, A2 spatial partitions are considered with sizes larger than 3 chromatin regions.

ii. Nucleoli partitions:

Following the same protocol as in Case 2 for speckle partitions, we first identify chromatin expected to have high nucleoli association. These regions are identified as those previously reported nucleoli associated domain (NAD)[53] regions and nucleolus organizing regions (NOR, on short arms of chromosomes 13, 14, 15, 21, and 22). Using these regions, we generate CINs in each structure of the population. Approximate nucleoli locations are then identified as the center of mass of the resulting spatial

partitions identified by Markov clustering of the CINs. Only top 25% largest spatial partitions are used as predicted nucleoli. For NOR regions, we use the first 25 restrained 200-kb regions that are closest in sequence to NOR regions in these five chromosomes, as NOR regions do not have Hi-C data and they are not restrained during the modeling protocol.

Properties of partitions

Size of partitions: The size of a spatial partition is calculated as $0.2\ x\ N$ Mb where *N* is the number of nodes in the partition that represents a 0.2 Mb region.

Fraction of inter-chromosomal edges (contacts): For each spatial partition, the inter-chromosomal edge fraction (ICEF) is calculated as:

$$ICEF = \frac{E_{inter}}{E_{intra} + E_{inter}}$$

where E_{intra} and E_{inter} are number of intra- and inter- edges in the partition, respectively.

Structural features

Unless otherwise noted, mean values of structural features for each genomic region are calculated from 2 copies and 10,000 structures (total 20,000 configurations) in the following structural feature calculations.

- **Mean radial position (RAD, #1)**

Radial position of a chromatin region i in structure m is calculated as:

$$r_{i,m} = \frac{d_{i,m}}{R^{nuc}}$$

where $d_{i,s}$ is the distance of *i* to the nuclear center, and R^{nuc} is the nucleus radius which is 5 μm. $r_{i,s} = 0$ means the region *i* is at the nuclear center while $r_{i,s} = 1$ means it is located at the nuclear surface.

Other radial position related analyses

i. *Overlap of subcompartment borders and large radial position transitions:* To identify regions coinciding with large transitions in radial positions, we first calculate each region's gradient in radial position from their average radial position profiles. Peaks and valleys in the gradient profile coincide with the regions of large radial transitions in the chromosome and are identified with the *detect_peaks* python package[57]. We obtain 1408 regions with large radial transitions with minimum peak height

(mph) set to 0.01 (the gradient values range between -0.06 – 0.05.) to filter out regions with minimal radial transitions. We then check if these identified regions coincide with the subcompartment borders, i.e. where two neighboring chromatin regions are in different subcompartments. We determine an overlap if there is a subcompartment border within a 1-Mb window of a given identified region with a large radial transition.

ii. *Shell analysis:* To map the preferred positions of 200-kb regions in the nucleus, we divide the nuclear volume of each model into 5 concentric shells $L = \{L_1, L_2, L_3, L_4, L_5\}$ so that each shell contains the same amount of chromatin in each single structure. We then calculate the probability of a subcompartment s to be in any shell from L:

$$P_{s,L_k} = \frac{1}{M} \sum_{m=1}^{M} \frac{N_{s,L_k,m}}{N_s}$$

where $N_{s,L_k,m}$ is the number of regions from subcompartment s in shell L_k in structure m, N_s is the total number of regions in subcompartment s, and M is the total number of structures.

iii. *Comparison with GP-seq:* GP-Seq scores[23] are rescaled to have values between 0 – 1, where scores 0 and 1 correspond to a chromatin region being at the nuclear lamina and nuclear center, respectively[23]. Average radial positions extracted from our structures vary between 0.48 – 0.94 with higher values corresponding to proximity to nuclear lamina. For comparison with GP-Seq, we subtract the average radial positions from 1 and then rescale the values to be between 0 – 1.

iv. *Average radial positions of regions from different replication phases:* Genomic regions are divided into 6 groups (G1b, S1, S2, S3, S4, G2) based on their mapped replication phases[36]. For each group, the distribution of the average radial positions is then determined from the structure population.

- **Local chromatin fiber decompaction (RG, #2)**

<u>Radius of gyration of chromatin fiber</u>

The local compaction of the chromatin fiber at the location of a given locus is estimated by the radius of gyration (RG) for a 1 Mb region centered at the locus (i.e. comprising +500kb up- and 500 kb downstream of the given locus). To estimate the RG values along an entire chromosome we use a sliding window approach over all chromatin regions in a chromosome.

The RG for a 1 Mb region centered at locus *i* in structure *m*, is calculated as:

$$RG_{i,m} = \sum_{j=1}^{N} {d_j}^2$$

where N is the number of chromatin regions in the 1-Mb window, and d_j is the distance between the chromatin region j to the center of mass of the 1-Mb region.

Other RG related analysis

i. *TAD border detection:* To investigate if chromatin regions with maxima in RG profiles coincide with TAD borders, we first identify peak regions in the average RG profiles with the *detect_peaks* python package[57]. 2068 peak regions are detected genome-wide with minimum peak distance (mpd) set to 3 (peaks must be at least 3 data points/600-kb apart from each other). We then check if these identified regions coincide with TAD borders detected by TopDom[58], HiCseg[59], InsulationScore[60], and TADbit[61]. We determine an overlap if there is a TAD border within ±200-kb window of a peak region.

ii. *RG peak frequency:* Peak regions in the RG profiles are detected in each individual structure using *detect_peaks* python package[57] with same parameters as in the previous section. The RG peak frequency (PF) of a region i is then calculated as:

$$PF_I = \frac{n_i + n_{i'}}{2M}$$

where n_i and $n_{i'}$ are the number of structures in which region i and its homologus copy has an RG peak, and M is the number of genome structures in the population.

- **Mean gene-speckle and gene-nucleolus distances (SpD, NuD, #3,4)**

For each 200-kb region, the closest speckle partition (or nucleolus partition) in each single structure is identified and the center-to-center distance is calculated (from the center of the region to the geometric center of the partition). The distances across the population are then averaged for each region to calculate mean speckle (or nucleolus) distances.

Other related analysis

Speckle distance heatmaps: A speckle distance heatmap for a chromosome visualizes, for a given chromatin region, the speckle distance variability across the population of models. For each copy of a chromatin region, the distance to the nearest predicted speckle is calculated in each structure of the population. These distances (20,000 distances total due to 2 copies and 10,000 structures) are ranked from lowest to highest values and plotted along a column of the speckle distance heatmap and color coded according to the distance. Colors range from low distance (red) to large distances (blue).

- **Cell-to-cell variability of features (δ_{RAD}, δ_{RG}, δ_{SpD}, δ_{NuD}, #5-8)**

Cell-to-cell variability of any structural feature (δ_I^{RAD} for radial positions, δ_I^{SpD} speckle distances, δ_I^{NuD} nucleoli distances, and δ_I^{RG} local decompaction) for a chromatin region I is calculated as:

$$\delta_I^F = log_2 \frac{\sigma_I^F}{\overline{\sigma^F}}$$

where σ_I^F is the standard deviation of the values for structure feature F calculated from both homologous copies of the region across all 10,000 genome structures in the population; $\overline{\sigma^F}$ is the mean standard deviation of the feature value calculated from all regions within the same chromosome of region I. Positive δ_I^F values ($\delta_i^F > 0$) result from high cell-to-cell variability of the feature (e.g. radial position); whereas negative values ($\delta_i^F < 0$) indicate low variability.

- **Interior localization frequency (ILF, #9)**

For a given 200-kb region, the interior localization frequency (ILF) is calculated as:

$$ILF_I = \frac{n_{r<0.5}}{M}$$

where $n_{r<0.5}$ is the number of structures where either copy of the region I has a radial position lower than 0.5, and M is the total number of structures which is 10,000 in our population.

- **Nuclear-body association frequencies (SAF, LAF, NAF, #10-12)**

For a given 200-kb region, the association frequency to nuclear bodies (SAF, LAF, and NAF for speckle, lamina, and nucleoli association frequencies, respectively) are calculated as:

$$SAF(or\ LAF\ or\ NAF)_I = \frac{n_{d_i<d_t} + n_{d_{i'}<d_t}}{2M}$$

where M is the number of structures in the population (2 homologous copies of each chromosome are present per structure); $n_{d_i<d_t}$ and $n_{d_{i'}<d_t}$ are the number of structures, in which region i and its homologous copy i' have a distance to the nuclear body of interest (NB) smaller than the association threshold, d_t. The d_ts are set to 500 nm, $0.35xR_{nuc}$, and 1000 nm for SAF, LAF, and NAF, respectively. We try different distance thresholds, and the select thresholds resulted in the best correlations with experimental data. For SAF and NAF calculations, we use the predicted speckle and nucleolus partitions to calculate distances (see *Identifying spatial partitions via Markov clustering*). For LAF, we use the direct

distances of regions to the nuclear envelope. For all association frequency calculations, we calculate distances from the surface of the region to the center-of-mass of the partition or to the surface of the nuclear envelope.

Other related analyses

i. *Predicting laminB1 DamID signals using LAF:* The predicted laminaDamID signal of region I is calculated as:

$$predicted\ laminaDamID\ signal_I = log_2\left(\frac{LAF_I}{\overline{LAF}}\right)$$

where $\overline{LAF}$ is the mean lamina association frequency calculated from all regions in the genome.

ii. *Comparison with imaging data:* We compare our SAF, LAF and NAF values with imaging data[19]. To calculate association frequencies from imaging and models, we use different distance thresholds (250, 500, 750, 1000 nm distance thresholds for SAF and LAF when calculated from imaging or models, and additional thresholds of 1250, 1500, 1750, 2000 nm for LAF when calculated from models) to define an association to the nuclear body of interest. We find that the best correlations are obtained when the following distance thresholds are used:

- SAF: 500 nm for imaging, 750 nm for models

- NAF: 1000 nm for imaging, 1000 nm for models

- LAF: 1000 nm for imaging, 2000 nm for models

For SAF comparisons, we use the predicted speckle partitions from interior regions (Case 2 for speckle partitions in *Identifying spatial partitions via Markov clustering*).

- **TSA-seq (S-TSA, L-TSA, N-TSA, #13-15)**

To predict TSA-seq signals for speckle, nucleoli, and lamina from our models, we use the following equation:

$$sig_i = \frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{L} e^{-R_0\|d_{il}\|}$$

where M is the number of models, L is the number of predicted speckle locations in structure m, d_{il} is the distance between the region i and the predicted nuclear body location l, and R_0 is the estimated decay constant in the TSA-seq experiment[17] which is set to 4 in our calculations. The normalized TSA-seq signal for region i then becomes:

$$predicted\ TSAseq\ signal_i = log\left(\frac{sig_i}{\overline{sig}}\right)$$

where $\overline{sig}$ is the mean signal calculated from all regions in the genome. The predicted signal is then averaged over two copies for each region. The predicted speckle, and nucleoli partitions are used for distance calculations (see *Identifying spatial partitions via Markov clustering*). For lamina TSA-seq, we use direct distances of each 200-kb chromatin region to the nuclear surface in each structure, which is calculated as $(1 - r_m) \times R_{nuc}$ where r_m is the radial position of the 200-kb region in structure m and R_{nuc} is the nucleus radius which is set to 5 μm.

<u>Other related analysis:</u>

i. *Predicting SON TSA-seq signals using only cis relationships in folded chromosomes:*

To identify contributions of cis interactions in SON TSA-seq signals, speckle locations are defined by the geometric center of consecutive A1 sequence blocks formed by more than 1 A1 chromatin region (instead of the geometric center of A1 spatial partitions, which can be formed by both cis and trans chromosomal interactions). For single A1 regions, the bead center location is used instead. For each chromatin region, we then calculate its spatial distances to these predicted speckle locations in the folded chromosome, which are used to predict the resulting TSA-seq signals from cis interactions only.

ii. *Predicting SON TSA-seq signals using only cis relationships in random conformations*:

We also repeat the same calculations as defined in the previous section, but instead of the folded chromosomes, use models with random chain configurations, generated without Hi-C data (i.e. only chain connectivity and excluded volume). TSA-seq data is calculated accordingly from the corresponding distances based on the random polymer chain configurations.

iii. *Predicting SON TSA-seq signals using speckle distances based on A1 sequence locations:*

Speckle locations are approximated by the sequence positions of A1 regions, either as median sequence position for a block of consecutive A1 chromatin regions or the sequence positions of individual A1 regions, if their neighboring regions are not part of the A1 subcompartment. The distance d_{ij}^{seq} between a chromatin region *i* and speckle position *j,* separated in sequence by *n* chromatin regions, is then defined as $d_{ij}^{seq} = 2n \times R^{ex}$, where $R^{ex} = 118\ nm$ is the excluded volume radius of a chromatin region in the models (see *Genome representation*). These distances are then used to predict SON TSA-seq signals as defined above.

iv. *Histone modification histograms based on predicted SON TSA-seq deciles:*

Following the procedure described in ref[17], we divide the 200-kb chromatin regions in our models into 10 decile groups based on their predicted SON TSA-seq signals; deciles 1 and 10 contain regions with the

lowest and highest 10% predicted TSA-seq signals, respectively. We then count the number of mapped peaks of H3K27me3, H3K4me3, and H3K9ac as well as the number of A1, A2, A1+A2 regions in each decile, and calculate the fraction of histone modification peaks or A1/A2 regions accrued in each decile. For mapping histone modification peaks to 200-kb bins to match our models' resolution, see *Mapping experimental data to models* in *Supplementary Information*. Same histograms using experimental TSA-seq deciles are re-generated from Fig. 8 in ref[17] using WebPlotDigitizer[62].

- **Mean inter-chromosomal neighborhood probability (ICP, #16)**

For each target chromatin region i, we define the neighborhood $\{j\}$ if the center-to-center distances of other regions $\{j\}$ to the target region are smaller than 500 nm, which can be expressed as a set; $Ne_i = \{j: j \neq i, d_{ij} < 500\ nm\}$. Inter-chromosomal neighborhood probability (ICP) is then calculated as:

$$ICP_I = \frac{1}{2M} \sum_{m=1}^{M} \sum_{i=1}^{2} \frac{n_{inter}(m,i)}{n_{inter}(m,i) + n_{intra}(m,i)}$$

where M is the number of structures, $n_{intra}(m,i)$ and $n_{inter}(m,i)$ are the number of intra- and inter-chromosomal regions in the set Ne_i in structure m for haploid region i.

- **Median trans A/B ratio (#17)**

For each chromatin region i, we define the trans neighborhood $\{j\}$ if the center-to-center distances of other regions from other chromosomes to itself are smaller than 500 nm, which can be expressed as a set; $Ne_i^t = \{j: chrom_i \neq chrom_j, d_{ij} < 500\ nm\}$. Trans A/B ratio is then calculated as:

$$trans\ AB\ ratio_i = \frac{n_A^t}{n_B^t}$$

where n_A^t and n_B^t are the number of trans A and B regions in the set Ne_i^t for haploid region i. The median of the trans A/B ratios for a region is then calculated from all the trans A/B ratios of the homologous copies of the region observed in all the structures of the population. The values are then rescaled to have values between 0 – 1.

Comparison of gene expression with structural features

- **Transcription frequency**

Transcription frequency (TRF) of each gene in the scRNA-seq data is defined as the fraction of cells in the population of cells, where the gene has non-zero mRNA transcription counts in the scRNA-seq data[24]. TRF is also calculated from the recently published nascent RNA-MERFISH imaging data as the fraction of cells where the gene is transcribed (transcription: on) in the population of imaged cells[19].

- **Gene expression heatmaps**

Gene expression heatmaps for each chromosome visualize the variability of mRNA counts (the expression levels) for each gene in a population of cells[24]. For each chromatin region, the observed mRNA count in each cell of the population of models is ranked from highest to lowest values and plotted along a column. Colors ranged from high mRNA counts (red) to 0 (dark blue).

- **ROC curve for assessing performance to classify lowly or highly expressed genes**

We first identify the top 10% (T10) and the bottom 10% (B10) genes with the highest and the lowest total non-zero mRNA counts (i.e. gene expression values) in the scRNA-seq data[24]. Several structural features (mean radial positions, ILF, mean speckle distances, SAF, variability of radial positions and speckle distances) are then calculated for all chromatin regions mapped to T10 genes and B10 genes.

To determine the most informative structural features for distinguishing T10 genes from B10 genes, we perform receiver operator characteristic (ROC) analysis. Specifically, for each structural feature, we define 10 threshold levels, equally separating the range of values for each structural feature. Then we determine how well the gene in the T10 and B10 groups are separated by each threshold value by calculating the corresponding number of true positives/negatives (TP, TN) and false positive/negatives (FP, FN).

For each structural featue f and for each threshold level, t, the true positive rate (TPR) and false positive rates (FPR) are then calculated as

$$TPR_{t,f} = \frac{TP}{TP + FN}$$

$$FPR_{t,f} = 1 - \frac{TN}{FP + TN}$$

The ROC curves are then plotted for each feature using TPR/FPR values.

Other structural analyses

- **Experimental GRO-seq and TSA-seq data analysis**

Averaging TSA-seq and GRO-seq signals in concentric shells around subcompartment partitions:

To quantify average TSA-seq[17] and GRO-seq[41] signals for chromatin with respect to the distance to spatial partition centers of each subcompartment, the nuclear volume around a spatial partition center is divided into concentric shells, with each consecutive shell radius increasing by 200 nm. The signals are then averaged over concentric shells around partition centers as follows: In each individual genome structure, the signals of chromatin located in the same shell volume is averaged, irrespective of the chromatin's subcompartment assignment. The average signal per shell are further averaged over all partition centers in the same subcompartment and over all structures of the population. Note that this measure only relies on the geometric position of a partition center and the folded genome (i.e. calculates average gene expression from all chromatin in a shell, independent of subcompartment annotations).

- **Neighborhood composition**

The neighborhood composition (NeC) shows how frequent chromatin regions in different subcompartments are in spatial proximity to regions of a specific subcompartment. The average percentage of subcompartment Q in the neighborhood composition of subcompartment S in the population is calculated as:

$$NeC_{SQ} = \frac{1}{MN_S} \sum_{m=1}^{M} \sum_{j=1}^{N_S} \frac{n_{Q,m,j}}{|N_{m,i}|} x100$$

where M is the number of structures in the population, N_S is the number of 200-kb regions belonging to subcompartment S, $\{N_{m,i}\}$ is the set of 200-kb chromatin regions in the neighborhood of the region i in structure m, and $n_{Q,m,i}$ is the number of chromatin regions from subcompartment Q in the set $\{N_{m,i}\}$. We define the neighborhood of i in structure m as $N_{m,i} = \{j: j \neq i, d_{ij} < 500\ nm\}$, which contains the list of all chromatin regions with less than 500 nm center-to-center distance (d_{ij}) to chromatin region i.

The neighborhood composition enrichment (NeCE) of subcompartment Q in the neighborhood of subcompartment S is calculated as:

$$NeCE_{SQ} = \frac{NeC_{SQ}}{\frac{1}{5}\sum_{T\in\{A1,A2,B1,B2,B3\}} NeC_{TQ}}$$

where NeC_{SQ} is the neighborhood composition percentage calculated for subcompartment Q in the neighborhood of subcompartment S and the denominator is the average percentage of subcompartment Q observed in the neighborhood of all subcompartments. Values greater than 1 ($NeCE_{SQ} > 1$) indicate that subcompartment Q is enriched in the neighborhood of subcompartment S, whereas values lower than 1($NeCE_{SQ} < 1$) show depletion of Q around S.

- **Structural feature enrichment heatmap**

To identify structural feature enrichments for chromatin in different groups (subcompartments, TSA-seq deciles, superenhancers, enhancers, replication phases, and T10/B10 genes), we first normalize each feature value to range between 0 and 1. We then calculate the enrichment of a structural feature f, for group g as:

$$enrichment_{g,f} = log_2 \frac{\frac{1}{N_g}\sum_{c=1}^{N_g} f_c}{\bar{f_r}}$$

where N_g is the number of 200-kb chromatin regions in group g, f_c is the structure feature value for chromatin region c. For $\bar{f_r}$, we first randomly select the same number (N_g) of regions in the genome and calculate the average feature value, and repeat this step 1000 times. We then take the average of 1000 different average feature values calculated from randomly selected regions.

For visualization purposes, we reverse the ranges of radial positions, mean-speckle, and mean-nucleoli distances in the enrichment heatmaps, so lower values would be indicated with red.

- **K-means clustering of A and B compartments**

For clustering, we first normalize all 17 structural features using log_2-transformation. We then perform K-means clustering using all transformed features for A and B subcompartments separately. We use scikit-learn python package to perform K-means clustering[63] and set the *n_clusters* parameter to 2 for A and 3 for B compartments. Clusters are then compared with actual subcompartment assignments to compute clustering accuracy. The highest prediction accuracies are obtained when clustering is performed with a subset of structural features for both A and B subcompartments. The used features in the clustering are cell-to-cell variability of radial positions, SAF, NAF, median trans A/B ratios for A, and cell-to-cell variability of radial positions and nucleoli distances, nucleoli TSA-seq, ICP, median trans A/B ratios for B subcompartment predictions, respectively.

- **Enrichment of histone marks in A2 regions with low and high variability:**

The enrichment/depletion of histone marks observed in highly variable (HV) A2 regions ($\delta_{RAD} > Q3$) is calculated as:

$$Enrichment_{A2,HV} = \frac{\frac{1}{N_{HV}} \sum_{i=1}^{N_{HV}} sig_i}{\frac{1}{N_{rest}} \sum_{i=1}^{N_{rest}} sig_i}$$

where N_{HV} and N_{rest} are the number of A2 HV regions and the rest of the A2 regions, respectively, and sig_i is the histone modification signal for region i. The same equation is also used to calculate the enrichment/depletion of histone marks observed in A2 regions with low variability (LV, $\delta_{RAD} < Q1$) compared to the rest of the A2 regions. For each enrichment calculation (HV or LV), A2 regions are divided into two groups: 1) HV and rest of A2; 2) LV and rest of A2. Note that, the rest of A2 regions in those groups are not the same.

For comparison, the enrichment/depletion of histone marks observed in same number of randomly selected regions (as in N_{HV} or N_{LV}) is calculated as:

$$Enrichment_{A2,R} = \frac{\frac{1}{N_R} \sum_{i=1}^{N_R} sig_i}{\frac{1}{N_{rest}} \sum_{i=1}^{N_{rest}} sig_i}$$

where N_R and N_{rest} are the number of randomly selected A2 regions and the rest of the A2 regions (used in the HV or LV enrichment calculations), respectively, and sig_i is the histone modification signal for region i. This calculation is repeated 1,000 times, and the average and the standard deviation of those 1,000 enrichment scores for randomly selected regions were used for comparison.

- **Comparison with 3D in situ hybridization (3D-FISH) data**

FISH probes are mapped to 200-kb chromatin regions in our models according to the highest overlap. Radial positions and pairwise distances for each mapped probe are determined in each structure in the population and compared to the radial positions and pair distances in FISH experiments. FISH and model radial positions are normalized by their maximum values. Intra-chromosomal distances in models are defined by their surface-to-surface distances of the corresponding probe regions (in both copies of the chromosome). Colocalization fraction of inter-chromosomal pairs are calculated as following: first the center-to-center distances of all possible probe pairs ($i - j,\ i - j',\ i' - j,\ i' - j'$ where i' and j' are the homologous copies of each 200-kb chromatin regions, i and j) are calculated in each structure. The minimum distance from all possible pairs in each structure is then used to calculate the fraction of models

in which both regions are colocalized. We assume a loci pair is colocalized in a structure if the calculated minimum distance in that structure is lower than 1 μm ($d_{min} < 1\ \mu m$).

Data visualization

CINs are visualized by Cytoscape[64]. 3D models and spatial partitions are visualized by using Chimera[65].

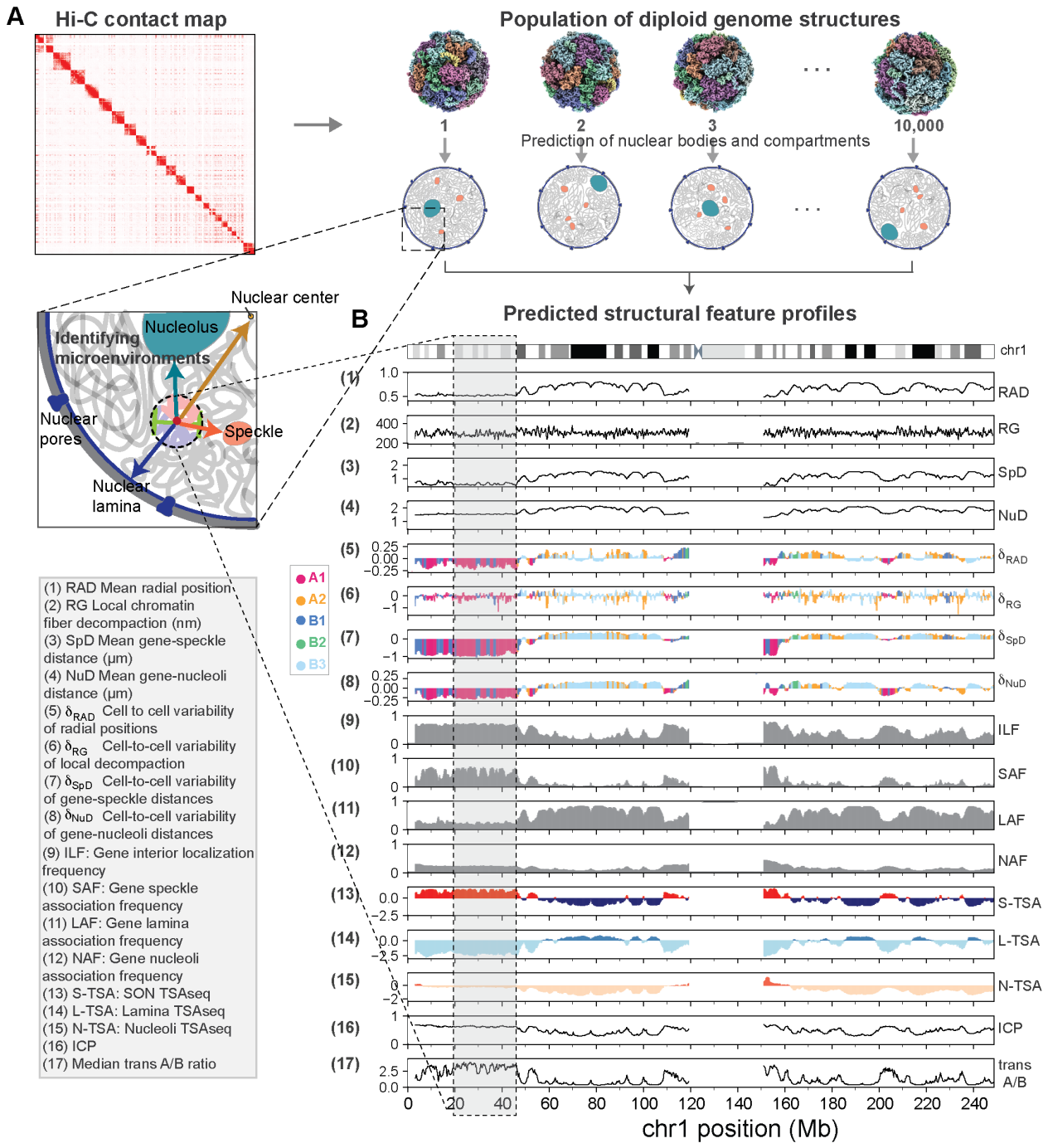


Fig. 1. Microenvironment and structural features. (**A**) Schematic depiction of our approach. A population of 10,000 genome structures is generated that is statistically consistent with the ensemble Hi-C data. Genome structures are used to predict the locations of nuclear speckles, nucleoli and the lamina associated compartment, which serve as reference points to describe the global genome organization and define structural features. (**B**) 17 structural features are calculated from the models that describe the nuclear microenvironment of each genomic region. Structure feature profiles for chromosome 1 are shown.

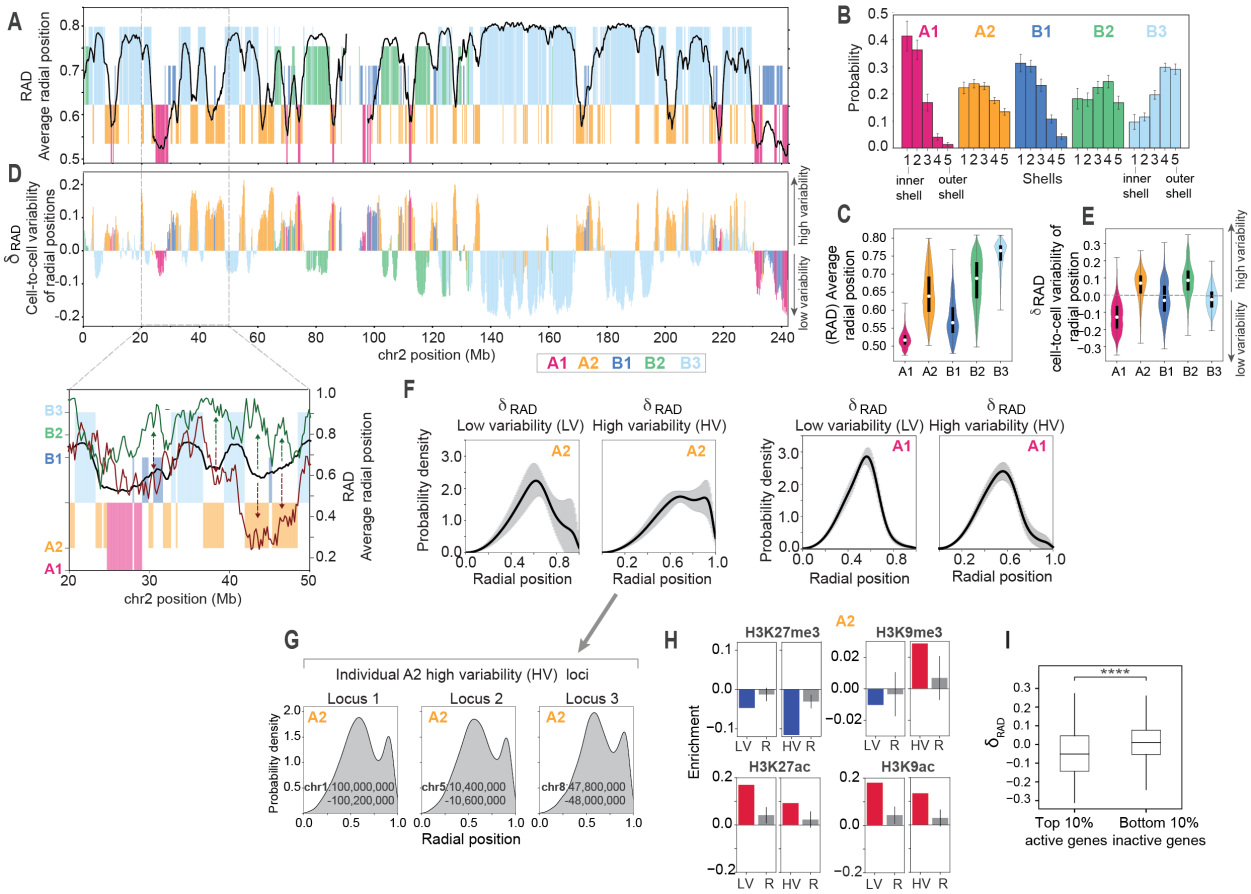


Fig. 2. Radial positions of chromatin and their cell-to-cell variability. (**A**) Average radial positions of chromatin regions in chromosome 2. Colored blocks indicate subcompartment assignments for each chromatin regions (A1: pink, A2: yellow, B1: dark blue, B2: green, B3: light blue) (**B**) Probabilities for chromatin region of a given subcompartment to be located in any of five concentric shells, each containing the same total amount of chromatin (*Methods*). Shell 1 is the most interior shell. Error bars show standard deviation. (**C**) Violin plots for the distributions of average radial positions for all chromatin regions in a subcompartment. White circles and black bars show the median value and the interquartile range (IQR: Q1 – Q3), respectively. (**D**) (upper panel) Cell-to-cell variability (δ^{RAD}) for chromatin regions in chromosome 2. $\delta_i^{RAD} = log_2(\sigma_i/\langle\sigma\rangle)$ with σ_i as the standard deviation of radial positions for chromatin region *i* in the structure population and $\langle\sigma\rangle$ as the average standard deviation for all chromatin regions within the same chromosome. Color-code is based on subcompartment annotations as in A. (lower panel) Zoomed-in radial position profiles for a 30 Mb region in chromosome 2. The black line shows the average radial positions, whereas green and maroon lines show radial positions in two different single structures. Arrows depict regions with high cell-to-cell variability. (**E**) Violin plots for distributions of cell-to-cell variabilities of radial positions for chromatin regions in different subcompartments. (**F**) Probability density distributions for radial positions of A2 regions with low (δ^{RAD} < Q1) and high (δ^{RAD} > Q3) cell-to-cell variability (left panel), and for radial positions of A1 regions with low (δ^{RAD} < Q1) and high (δ^{RAD} > Q3) cell-to-cell variability (right panel). Black lines indicate the average distribution, gray areas show the standard deviation calculated from all the regions within each group. (**G**) Radial distributions of three representative individual A2 regions with high-cell to cell variability observed in the structure population. (**H**) Enrichment of two active (H3K27ac, H3K9ac) and two inactive (H3K27me3, H3K9me3) histone marks in A2 loci with high and low variability compared to the rest of the A2 regions. Bars labeled with "R" show the enrichment in the randomly selected same number of A2 loci. Error bars in R bars show standard deviation calculated

from 1000 individual random selections. (**I**) Box plots for distributions of cell-to-cell variabilities of radial positions for chromatin regions where the top 10% highly transcribed genes and the bottom 10% genes with low transcriptional activity are located according to scRNA-seq data from Osorio *et al.*[24].

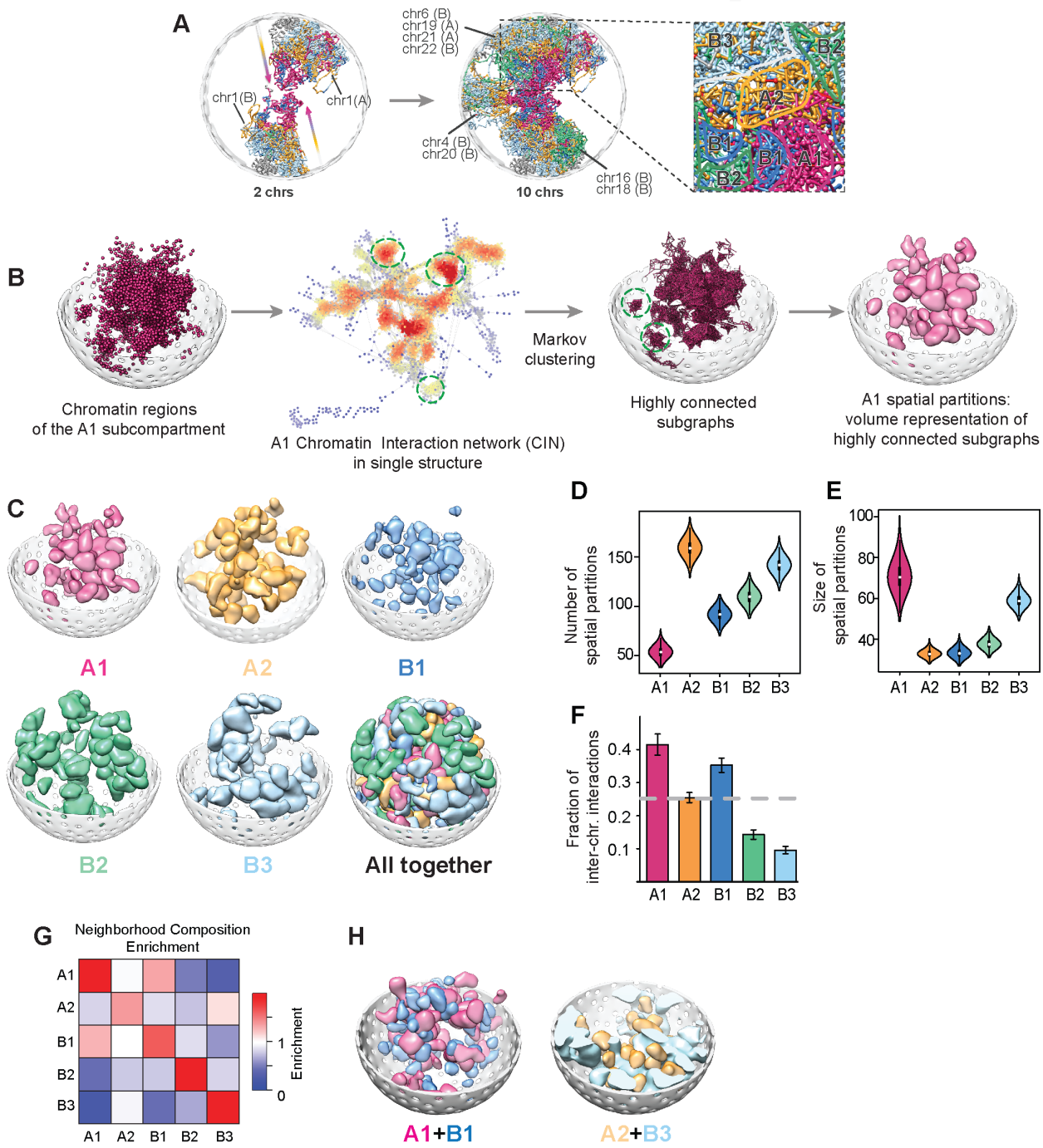


Fig. 3. Spatial partitions of subcompartments. (**A**) A representative genome structure showing chromosome folding patterns. For clarity both images show the same single structure with different numbers of chromosomes. Zoomed inset delineates regions that are primarily occupied by chromatin of the same subcompartment. Color-code indicates subcompartment annotations for each chromatin region (A1: pink, A2: yellow, B1: dark blue, B2: green, B3: light blue). (**B**) Procedure to identify spatial partitions of subcompartments: A chromatin interaction network (CIN) is generated from all chromatin regions in a given subcompartment for each structure in the population. Each node in the CIN represents a single

chromatin region connected by edges if the two regions are in physical contact in the 3D structure. Nodes are colored by their neighborhood connectivity (i.e. the average contacts formed by their neighbor nodes) ranging from low (blue) to high (red). Highly connected subgraphs are then identified by Markov Clustering algorithm of CINs (*Methods*) and visualized in the 3D structure (some shown in green dashed circles). The rightmost image illustrates the volume occupied by a spatial partition in a single genome structure. (**C**) Spatial partitions of subcompartments, shown by their occupied volume in the 3D structures. Only the 50 largest partitions (i.e. subgraphs with the largest numbers of nodes) are shown per subcompartment for clarity. (**D**) Distributions of the number of subcompartment partitions per genome structure. (**E**) Distributions of the average size (i.e. number of nodes) of subcompartment partitions. White circles and black bars in the violins show the median value and the interquartile range (IQR: Q1 – Q3), respectively. (**F**) Average fraction of inter-chromosomal edges in spatial partitions for each subcompartment. Error bars indicate standard deviations. Gray dashed line shows the average fraction for all partitions combined. (**G**) Neighborhood enrichment of chromatin in each subcompartment, defined as the ratio of (observed/expected) fraction of subcompartment chromatin in the immediate neighborhood (within 500 nm) of each chromatin region (*Methods*). **(H)** A representative structure showing examples of colocalizations of A1-B1 and A2-B3 partitions in the 3D space.

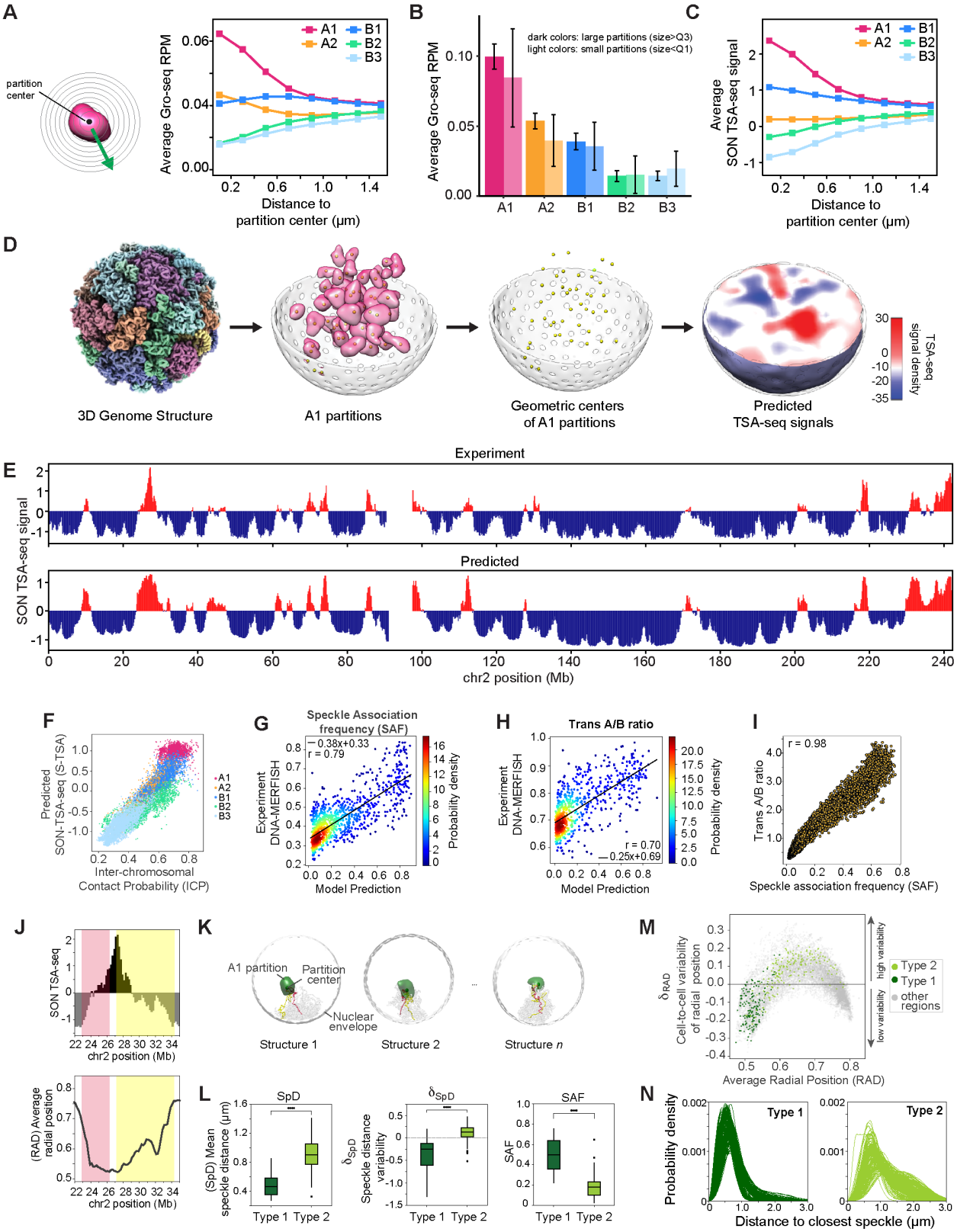

A
partition center
Average Gro-seq RPM
A1
A2
B1
B2
B3
Distance to partition center (μm)
B
dark colors: large partitions (size>Q3)
light colors: small partitions (size<Q1)
Average Gro-seq RPM
C
Average SON TSA-seq signal
Distance to partition center (μm)
D
3D Genome Structure
A1 partitions
Geometric centers of A1 partitions
Predicted TSA-seq signals
TSA-seq signal density
E
Experiment
Predicted
SON TSA-seq signal
chr2 position (Mb)
F
Predicted SON-TSA-seq (S-TSA)
Inter-chromosomal Contact Probability (ICP)
G
Speckle Association frequency (SAF)
y = 0.38x+0.33
r = 0.79
Experiment DNA-MERFISH
Model Prediction
Probability density
H
Trans A/B ratio
Experiment DNA-MERFISH
r = 0.70
y = 0.25x+0.69
Model Prediction
Probability density
I
r = 0.98
Trans A/B ratio
Speckle association frequency (SAF)
J
SON TSA-seq
chr2 position (Mb)
(RAD) Average radial position
K
A1 partition
Partition center
Nuclear envelope
Structure 1
Structure 2
Structure n
M
δRAD Cell-to-cell variability of radial position
high variability
low variability
Type 2
Type 1
other regions
Average Radial Position (RAD)
L
SpD
(SpD) Mean speckle distance (μm)
δSpD
δSpD Speckle distance variability
SAF
Type 1
Type 2
N
Type 1
Type 2
Probability density
Distance to closest speckle (μm)

Fig. 4. SON TSA-seq predictions using 3D structures. (A) Expression of genes with respect to their positions in spatial partitions. Average GRO-seq signal (RPM) of chromatin with respect to their 3D distance to spatial partition centers. The nuclear volume at each partition center is divided into concentric shells (left) and the nascent RNA expression levels (from GRO-seq experiments[41]) are averaged over all chromatin located in each concentric shell (irrespective of subcompartment annotations) (right). (**B**) Comparison of average GRO-seq signals for chromatin in large (size>Q3, dark colors) and small (size<Q1, light colors) spatial partitions for different subcompartments. Error bars represent standard deviation. (**C**) Average SON TSA-seq signal[17] of chromatin with respect to their distance to partition centers. The SON TSA-seq signals are averaged over all chromatin located in each concentric shell around partition centers (irrespective of subcompartment annotations). (**D**) The procedure for SON TSA-seq signal prediction from 3D models: A1 spatial partitions are identified in each structure of the population. The geometric centers of each A1 partition are used as approximate speckle locations and serve as point sources for the simulation of SON-TSA-produced tyramide free-radical diffusion[17] in single cell models. SON TSA-seq signals are then averaged over the population of all structures. (*Methods*). The rightmost image shows a cross section of the predicted TSA-seq signal density distribution in a single genome structure. (**E**) Comparison of the experimental and predicted SON TSA-seq profiles for chromosome 2 (Pearson corr. = 0.90, p=~0 for chromosome 2, 0.87, p=~0 for genome-wide prediction (see Extended Data Fig. 2a)). (**F**) Scatter plot of predicted SON-TSA-seq values and Inter-chromosomal contact probability (ICP), defined as the fraction of inter-chromosomal interactions among all interactions of a chromatin region. (Pearson corr. = 0.92, p=~0). (**G**) Comparison of predicted speckle association frequencies (SAF) in our models (*Methods*) with SAF determined with DNA-MERFISH experiments[19] for 1041 imaged loci. (**H**) Scatter plots of median trans A/B ratios predicted in our models (Methods) and determined by DNA-MERFISH experiments[19] for 724 imaged loci that belong to the same compartment in GM12878 and IMR-90 cells. (**I**) Scatter plots of predicted median trans A/B ratios as functions of predicted SAF for 1041 imaged by DNA-MERFISH experiments[19]. (**J**) Experimental SON TSA-seq data (top) and average radial position (bottom) for a ~11 Mb region of chromosome 2 showing a so-called TSA-seq trajectory, a a valley to peak to valley transition in the TSA-seq profile. (valley-to-peak: red region, peak-to-valley: yellow region). (**K**) Three representative structures showing folding patterns of the chromatin fiber for the ~11 Mb TSA-seq trajectory as in J. Shown are also the nuclear envelope, the closest A1 partition and the closest predicted speckle location. The chromatin fiber is color coded in red and yellow to represent corresponding regions shown in J. (**L**) Box plots of mean speckle distance (left panel), speckle distance variability (middle panel), and SAF (right panel) for regions where type 1 and type 2 TSA-seq peaks are located. (**M**) Scatter plot of average radial positions (RAD) against cell-to-cell variability of radial positions (δ_{RAD}) for of all genomic regions. Type 1 regions are shown in dark green color, type 2 regions in light green, and all other genomic regions in grey. (**N**) Distributions of gene-speckle distances for individual Type 1 loci (left) and Type 2 loci (right) in the population.

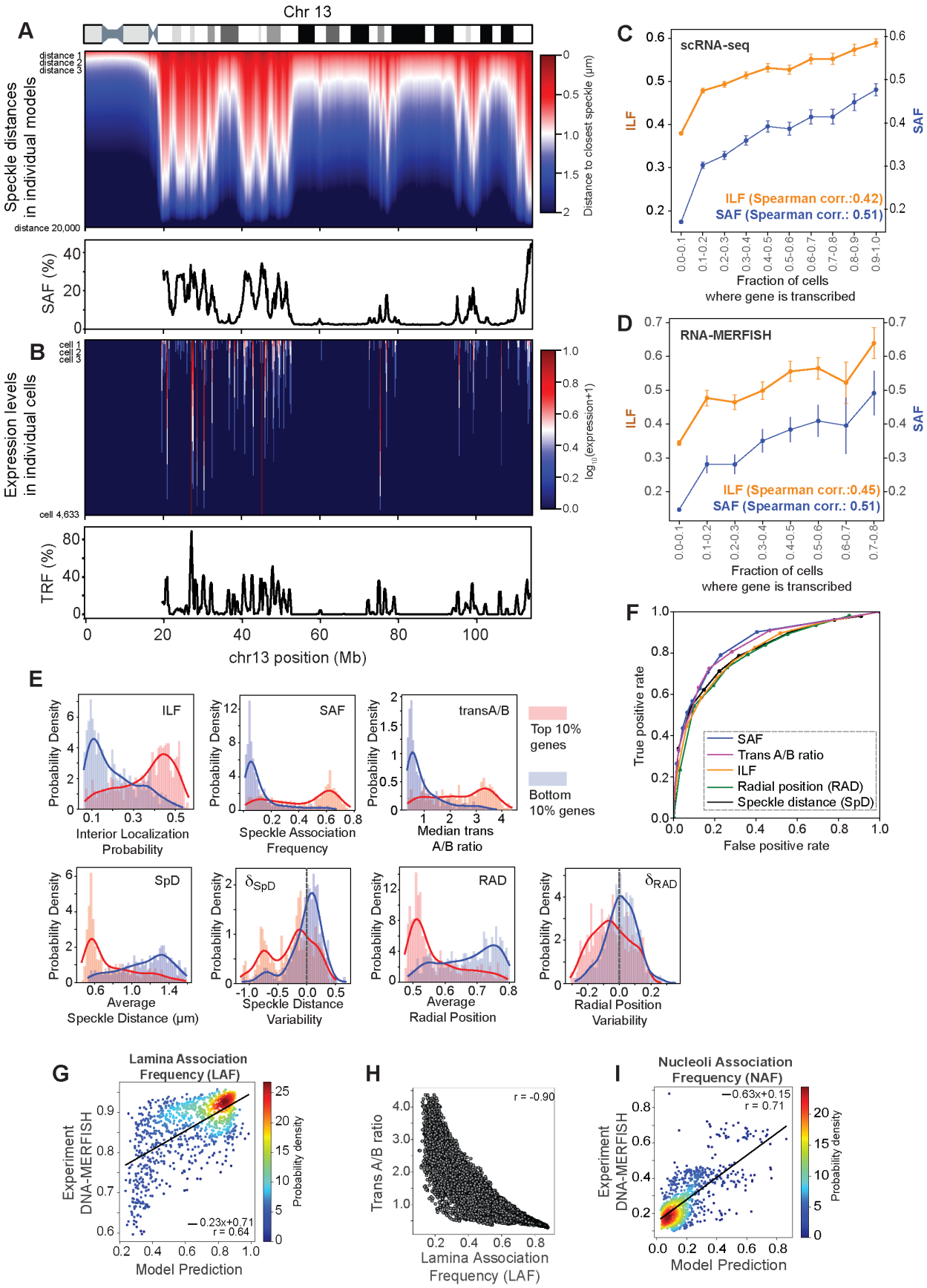

Chr 13
A
Speckle distances in individual models
distance 1
distance 2
distance 3
distance 20,000
Distance to closest speckle (µm)
SAF (%)
B
Expression levels in individual cells
cell 1
cell 2
cell 3
cell 4,633
log10(expression+1)
TRF (%)
chr13 position (Mb)
C
scRNA-seq
ILF
SAF
ILF (Spearman corr.:0.42)
SAF (Spearman corr.: 0.51)
Fraction of cells where gene is transcribed
D
RNA-MERFISH
ILF (Spearman corr.:0.45)
SAF (Spearman corr.: 0.51)
Fraction of cells where gene is transcribed
E
ILF
Interior Localization Probability
SAF
Speckle Association Frequency
transA/B
Median trans A/B ratio
Top 10% genes
Bottom 10% genes
SpD
Average Speckle Distance (µm)
δSpD
Speckle Distance Variability
RAD
Average Radial Position
δRAD
Radial Position Variability
Probability Density
F
True positive rate
False positive rate
SAF
Trans A/B ratio
ILF
Radial position (RAD)
Speckle distance (SpD)
G
Lamina Association Frequency (LAF)
Experiment DNA-MERFISH
Model Prediction
0.23x+0.71
r = 0.64
Probability density
H
Trans A/B ratio
Lamina Association Frequency (LAF)
r = -0.90
I
Nucleoli Association Frequency (NAF)
Experiment DNA-MERFISH
Model Prediction
0.63x+0.15
r = 0.71
Probability density

Fig. 5. Relationship between 3D chromatin structure and transcriptional activity. (**A**) (Top panel) Heatmap of gene speckle distances in chromosome 13 in all 10,000 structures. For a given gene, each column shows the gene-speckle distances in all 10,000 structures of the population. In each column, gene-speckle distances are sorted in ascending order from top-to-bottom, with short distances (dark red) to large distances (dark blue). (Bottom panel) Speckle association frequencies (SAF) for each chromatin region in chromosome 13. (**B**) (Top panel) Heatmap of single cell mRNA counts of genes in chromosome 13 in all 4,633 G1 cells measured by single cell RNA-seq experiment[24]. For a given gene, each column shows the observed mRNA transcript count in each cell of the population of cells. In each column, mRNA transcript counts are sorted in descending order from top-to-bottom, with high counts (dark red) to zero counts (dark blue). (Bottom panel) Transcription frequency (TRF) for each gene in chromosome 13. The TRF is the fraction of cells in the population of cells in which the gene has non-zero mRNA transcription counts in scRNA-seq data[24] (*Methods*). (**C**) Interior localization frequency (ILF) and SAF values for genes with different TRFs in scRNA-seq data[24]. (**D**) ILF and SAF values for genes with different TRFs from nascent RNA-MERFISH imaging[19]. Error bars show standard deviations of ILF and SAF values in each TRF range in C and D. (**E**) Distributions of several structural features for top 10% and bottom 10% genes based on their transcriptional activity obtained from scRNA-seq data[24]. Gray dashed line in cell-to-cell variability plot separates low (negative values) and high (positive values) levels of variability. (**F**) Receiver Operator Characteristic (ROC) curves for gene radial positions, speckle distances, ILF, SAF, and trans A/B ratios to distinguish actively transcribed genes with top 10% highest from 10% lowest transcription levels. SAF and trans A/B ratios are found to have the largest AUC (area under the curve) values; 0.85 and 0.84, respectively, compared to speckle distance (0.72), radial position (0.65), and ILF (0.81). (**G**) Comparison of predicted lamina association frequencies (LAF) in our models (*Methods*) with LAF determined from DNA-MERFISH experiments[19] for 1041 imaged loci. (**H**) Scatter plots of predicted median trans A/B ratios as functions of predicted LAF for 1041 imaged by DNA-MERFISH experiments[19]. (**I**) Comparison of predicted nucleoli association frequencies (NAF) in our models (*Methods*) with NAF determined from DNA-MERFISH experiments[19] for 1041 imaged loci.

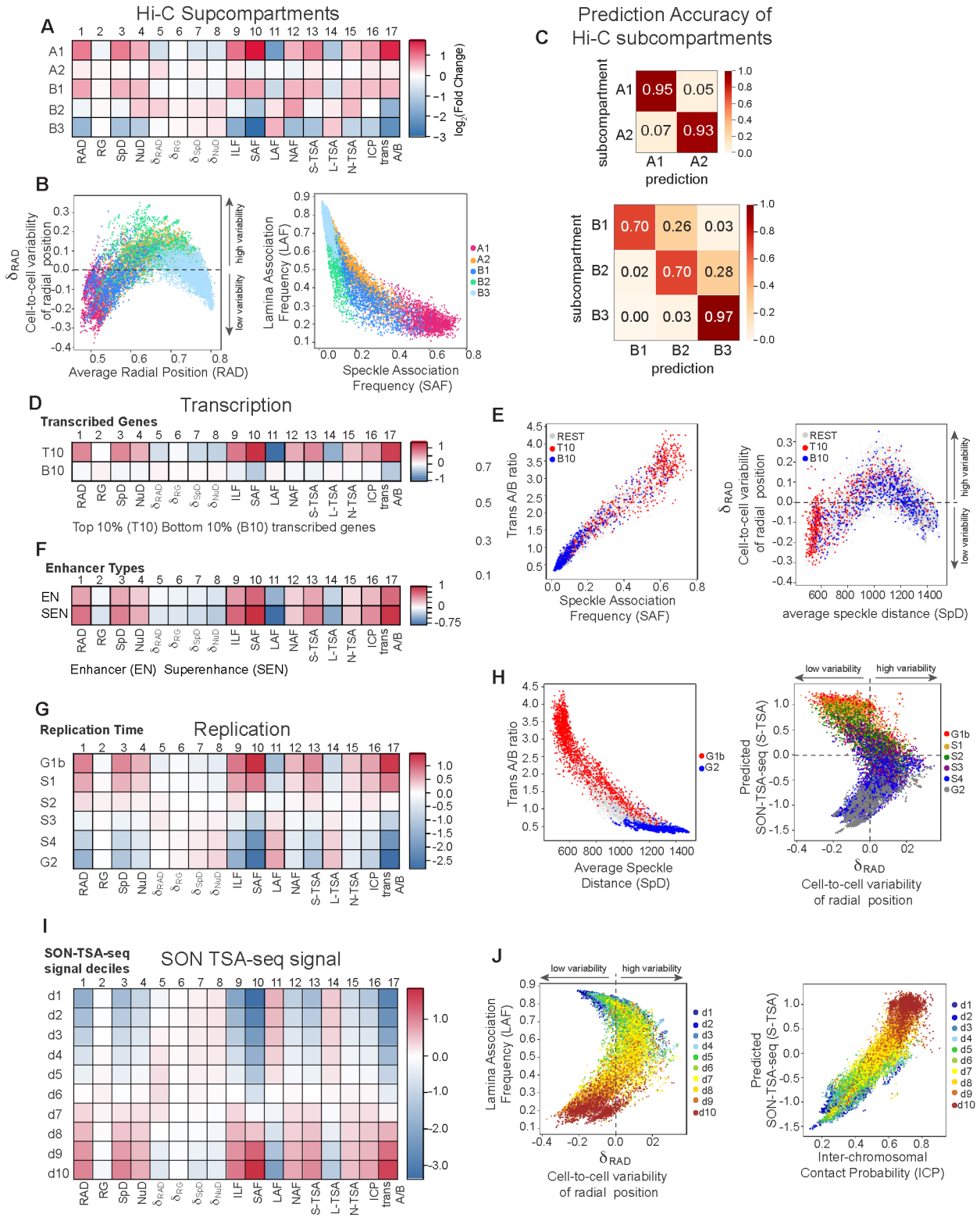
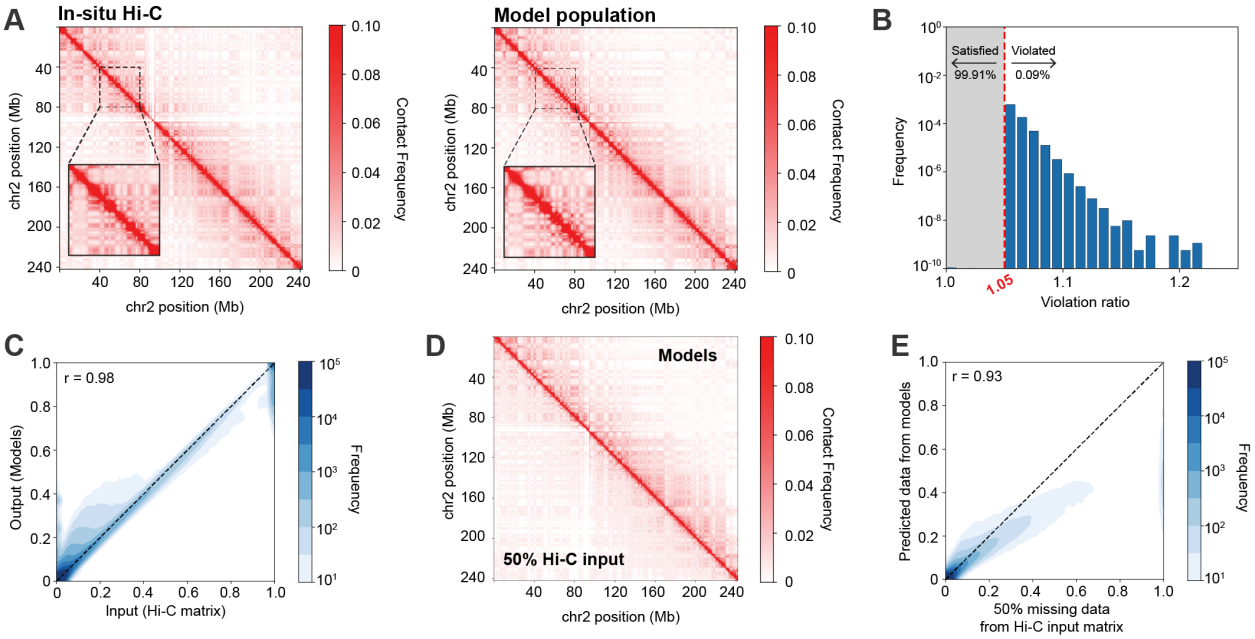


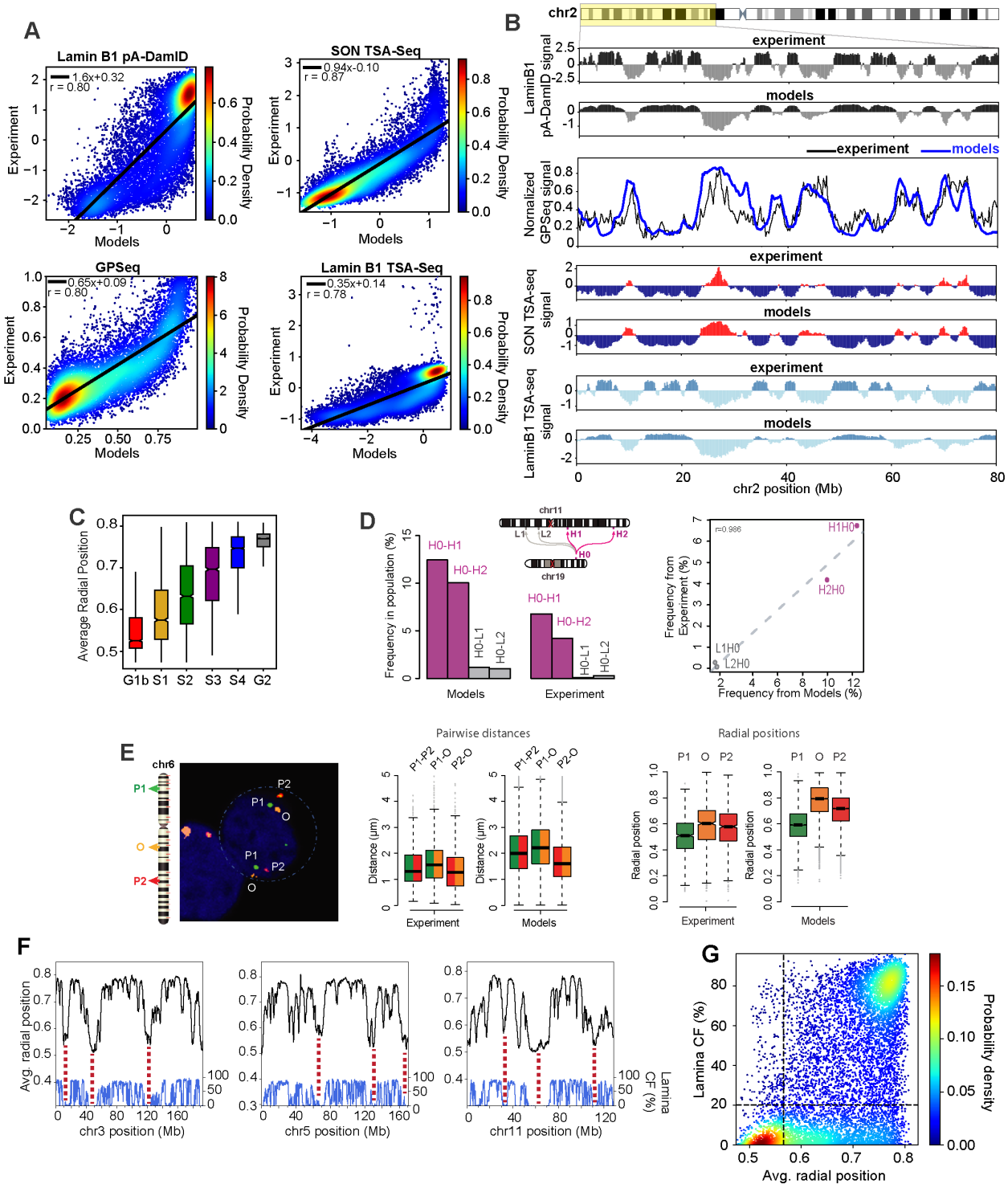
Fig. 6. Structural features of microenvironments. (**A**) Fold-change enrichment for each of the 17 structural features for chromatin in each subcompartment with respect to the genome-wide average (*Methods*). Note that speckle locations and all other features were calculated without knowledge of subcompartment assignments. (**B**) (Left panel) Scatter plot of δ^{RAD} (cell-to-cell variability of radial

positions) versus RAD (average radial positions) of chromatin regions. (Right panel) Scatter plot of Lamina Association Frequency (LAF) versus Speckle Association Frequency (SAF). Chromatin regions are color coded according to their subcompartment annotation. (**C**) Confusion matrices for the prediction of A1 and A2 (top) and B1, B2 and B3 (bottom) subcompartments using K-means clustering based on a gene's nuclear microenvironment (i.e., structural features) alone (*Methods*). (**D**) Fold-change enrichment for each of the 17 structural features for chromatin with 10% highest and 10% lowest transcription levels of transcribed genes (*Methods*). (**E**) (Left panel) Scatter plot of trans A/B ratio versus SAF. (Right panel) Scatter plot δ^{RAD} versus mean speckle distances (SpD). T10 genes are shown in red, B10 genes are shown in blue. (**F**) Fold-change enrichment for each of the 17 structural features for genes regulated by superenhancers (SE) and enhancers (E) (*Methods*). (**G**) Fold-change enrichment for each of the 17 structural features for genes replicated at different replication phases[36] (*Methods*). (**H**) (Left panel) Scatter plot of trans A/B ratio versus SpD (average speckle distance). Chromatin regions are color coded according to the time they are replicated: (red) earliest G1b and (blue) latest G2 phase. (Right panel) Scatter plot of predicted SON TSA-seq signals versus δ^{RAD}. Chromatin regions are color coded according to their replication timing[36]. (**I**) Fold-change enrichment of all structural features for chromatin regions divided by deciles of their experimental SON-TSA-seq values[17]. (**J**) (Left panel) Scatter plot of LAF versus δ^{RAD}. (Right panel) Scatter plot of predicted SON-TSA-seq signal versus ICP (inter-chromosomal Contact Probability). Chromatin regions are color coded according to their SON TSA-seq decile group in experiment. Gray dashed lines in B, E, H, G separate low (negative values) and high (positive values) levels of variability. The additional horizontal gray dashed line in H right panel also separates the positive and negative predicted SON TSA-seq signals.

Extended Data Figures and Tables

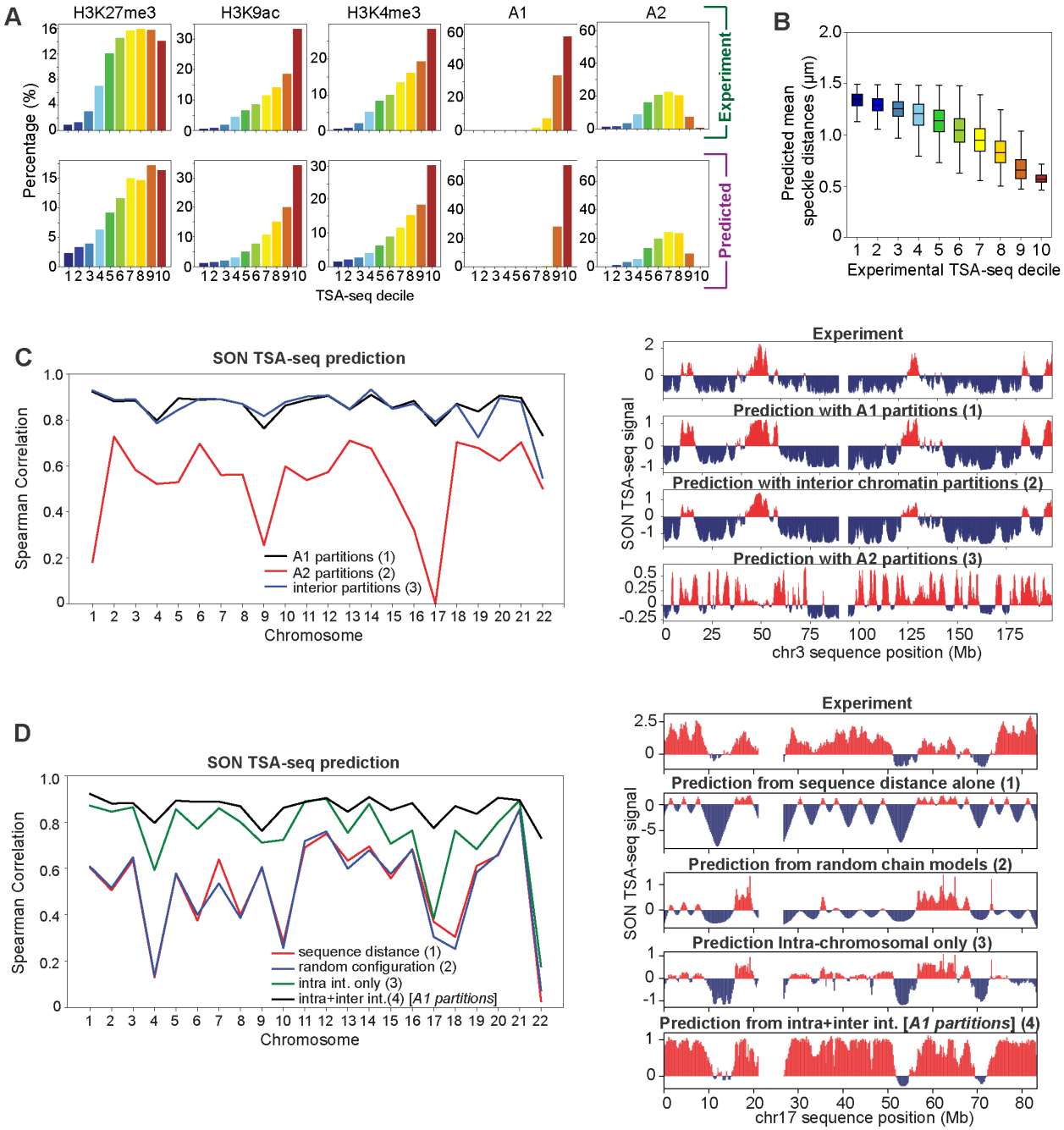


Extended Data Fig. 1. 3D chromatin structure modeling and assessment. (**A**) The Hi-C contact probability matrix (left) and the contact probability matrix calculated from the structure population (right) for chromosome 2. Zoomed-in heatmaps show the matrix between sequence position 40 – 80 Mb. (**B**) Histograms of restraint-violation ratio from the structure population. For a pair of constrained chromatin regions, violation ratio is defined as the ratio of the real distance over the expected distance (*Methods*). Violation ratio less than 1.05 is considered as satisfied and is not displayed in the histograms (99.9% of restraints fall in this category). (**C**) Density scatter plot comparing the contact probabilities from Hi-C data and structure population (Pearson's r = 0.98, p=~0). (**D**) The contact probability matrix for chromosome 2 showing the 50% randomly chosen dataset used as input (lower triangle) vs. the matrix generated from the structure population (upper triangle). (**E**) Density plot comparing the contact probabilities that are generated from Hi-C data and missing in the input and their predicted contact probabilities calculated from the structure population (Pearson's r = 0.93, p=~0).



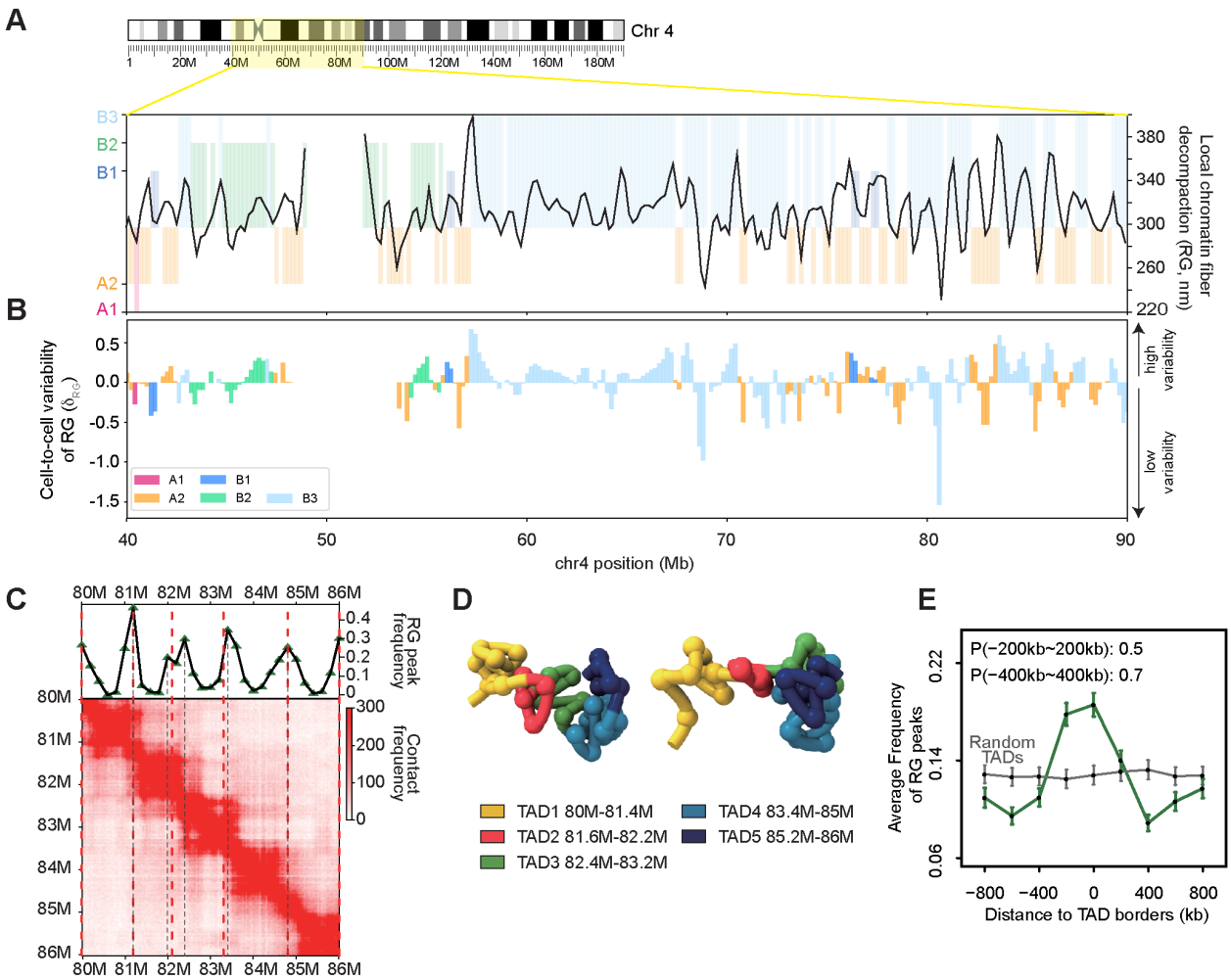
Extended Data Fig. 2. Assessment of 3D genome models with orthogonal data. (**A**) Scatter plots showing the comparison between experimental and predicted Lamin B1 pA-DamID[22] (top, left), SON TSA-seq[17] (top right), GP-seq[23] (bottom left), LaminB1 TSA-seq[17] (bottom right) (**B**) Profiles of experimental and predicted LaminB1 pA-DamID[22], GP-seq[23], SON TSA-seq[17], and LaminB1 TSA-seq[17] for the 0 – 80 Mb region in chromosome 2. (**C**) Average radial positions of chromatin in different replication phases[36]. (**D**) Comparison of the inter-chromosomal loci co-localization frequencies between the

observed occurrence in FISH experiments[37] and in the structure population. Histogram of co-localization events between a locus on chromosome 19 (labeled H0; an active domain) and any of the 4 loci on chromosome 11 (two inactive domains L1 & L2, and two active ones H1 & H2) from FISH experiments and predicted in the models. Scatter plot showing the co-localization frequencies from FISH experiments and the structure population (right panel). (**E**) A FISH image with three different probes at far-separating loci on chromosome 6 (left), the comparison of pair-wise distances of these loci in experiment and models (middle), and the comparison of their relative radial positions in experiment and models (right). (**F**) Average radial position profiles in chromosomes 3 (left), 5 (middle), and 11 (right). Also shown in blue are lamina CF from single cell lamin DamID experiments[35]. Valleys in the average radial position plots match well with low lamina CF regions (red dashed lines). (**G**) Density scatter plot of average radial positions of chromatin regions from the structure population against the lamina contact frequencies from single cell lamin DamID experiments in haploid KBM7 cell type (CF; DamID data from[35]). 93% of chromatin regions with the 25% lowest average radial positions show either no detectable or only occasional contact with lamina (CF < 20%). Vertical and horizontal black dashed lines show the 25th percentile average radial position and the 20% CF values, respectively.

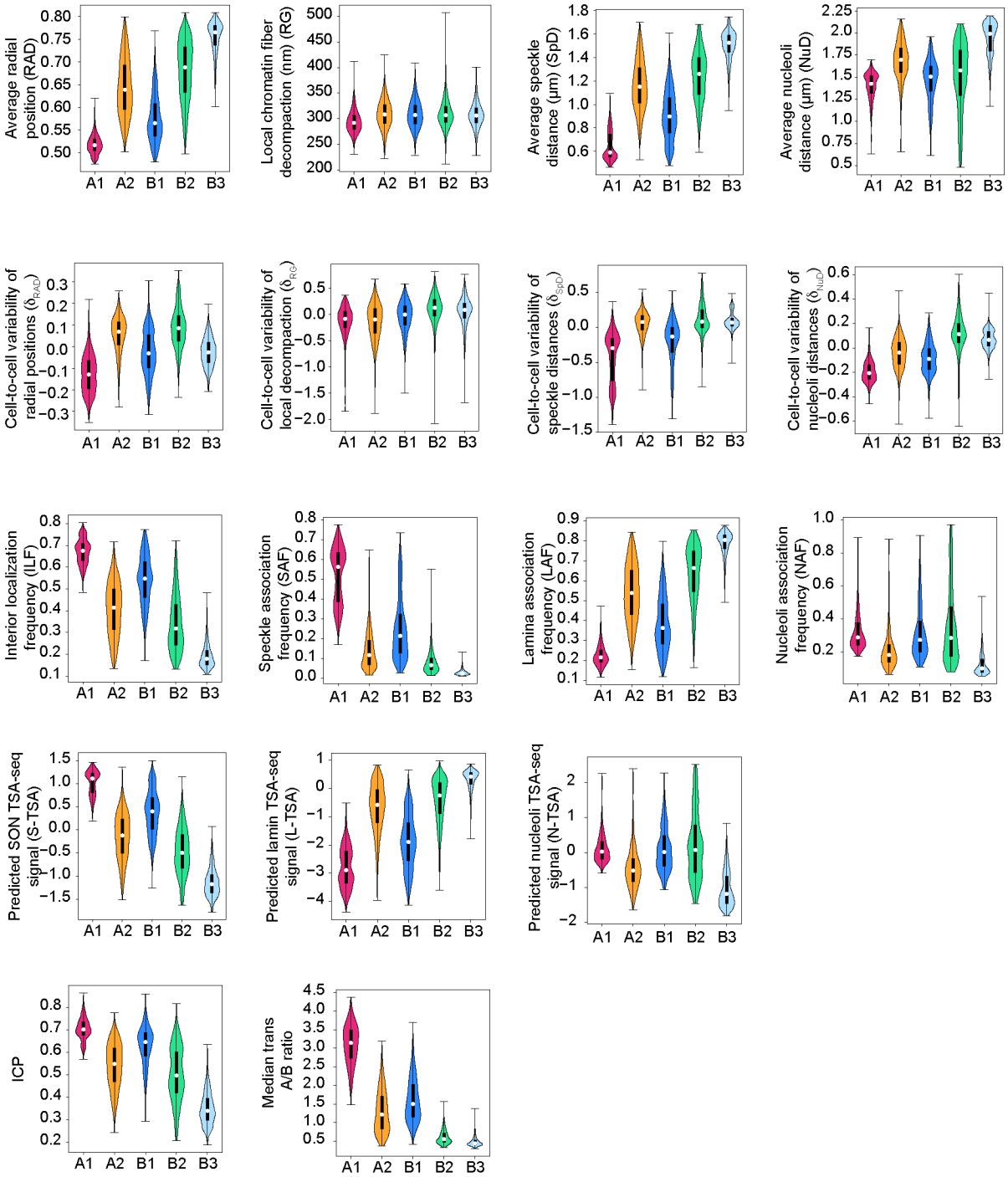


Extended Data Fig. 3. SON TSA-seq predictions using 3D models. (**A**) Fraction of mapped histone modifications peaks from ChIP-seq experiments as well as number of A1/A2 chromatin regions in chromatin divided into decile groups based on their TSA-seq signals (Methods). The upper panels show the analysis based on experimental TSA-seq data as done in ref[17]. Lower panels show the same analysis done based on predicted TSA-seq in our models. (**B**) Distributions of predicted mean distances to closest speckles (A1 partition centers) for chromatin regions in each experimental SON TSA-seq decile[17]. (**C**) (left panel) Spearman correlations between experimental SON TSA-seq data and predicted data in our models for each chromosome separately; (black line) predictions using A1 spatial partition centers, (red line) predictions using A2 spatial partition centers (red line), and (blue line) predictions using spatial partitions from chromatin with 10% lowest average radial positions in the population (*Methods*). (right

panels) Corresponding TSA-seq profiles of chromosome 3 for predicted and experimental data (Spearman correlations: 0.88, 0.89, 0.58 , respectively). (**D**) (left panel) Spearman correlations between the experimental SON TSA-seq signals and the predicted signals for each chromosome using different prediction methods (*Methods*): (red line) predictions using sequence distances to A1 clusters in sequence, (blue line) predicted using 3D distances to A1 partitions of in random chain chromosome territories; (green line) predicted using 3D distances to A1 regions in the same chromosome only; (black line) predictions using 3D distances to A1 partition centers using both intra-and interchromosomal relationships. (right panels) Corresponding TSA-seq profiles of chromosome 17 for predicted and experimental data (Spearman correlations: 0.37, 0.30, 0.38, 0.78, respectively).



Extended Data Fig. 4 Chromatin compaction and TAD borders. (**A**) Average radius of gyration (RG, i.e. local decompaction) profile for chromatin in the 40 – 90 Mb region of chromosome 4. The background is color coded by the subcompartment annotations of chromatin. (**B**) Cell-to-cell variability of RG values (δ^{RG}) in the structure population for the same chromatin regions. Negative values indicate regions with low RG variability. Bars are color coded by the subcompartment annotations of the corresponding chromatin regions. (**C**) (Top panel) RG peak frequencies (i.e., the fraction of models showing a RG maximum at a given position) for a 6-Mb region in chromosome 4 (80–86Mb). (Bottom panel) Hi-C contact frequency heat map for the same region showing TAD borders identified by TopDom[58] (red dashed lines). Regions with RG peak frequency maxima are shown with gray dashed lines, and either overlap or are very close to TAD borders identified by TopDom (red dashed lines). (**D**) Two representative structures showing chromatin folding patterns for chromatin regions in C. TAD identities are shown by color code. (**E**) Averaged RG peak frequencies for loci at TopDom TAD borders (green) compared to randomly selected loci (gray). In around 50% of structures, there is a RG peak in the immediate neighboring region of a TAD border (±200kb). In ~ 70% of structures there is a RG peak within a ±400kb range of a TAD border. Standard errors calculated from all TAD borders are shown with error bars.



Extended Data Fig. 5. Structural features of chromatin in different subcompartments. Violin plots for the distributions of 17 structural features calculated from the structure population for chromatin in different subcompartments. White circles and black bars in the violins show the median value and the interquartile range (IQR: Q1 – Q3), respectively.

Extended Data Table 1. Genome-wide correlations between experimental and predicted omics and imaging data. All p-values are ~0. Chromosome X is discarded from genome-wide correlation calculations in TSA-seq, DamID, and GPseq comparisons.

	Pearson's r	**Spearman's r**
SON TSA-seq[17] predictions with A1 partitions	0.87	0.89
SON TSA-seq[17] predictions with interior partitions	0.86	0.88
SON TSA-seq[17] predictions with A2 partitions	0.18	0.38
SON TSA-seq[17] predictions with A1 sequence distances	0.35	0.64
SON TSA-seq[17] predictions from random configurations (only-intra)	0.60	0.58
SON TSA-seq[17] predictions from folded chromosomes (only-intra)	0.73	0.79
LaminB1[17] TSA-seq predictions using direct distance to nuclear envelope	0.78	0.81
LaminB1 pA-DamID[22]	0.80	0.79
GP-seq[23]	0.80	0.79
SAF[19] using A1 partitions	0.77	0.73
SAF[19] using interior partitions	0.79	0.74
LAF[19]	0.64	0.58
NAF[19]	0.71	0.63
Median trans A/B ratio[19]	0.70	0.67

Extended Data Table 2. Properties of subcompartment interaction networks and spatial partitions. Population averages of features for chromatin interaction networks (CIN) and spatial partitions of chromatin in different subcompartments (Methods).

CIN/Partition Features	A1	A2	B1	B2	B3
Average neighborhood connectivity in CINs	25.92	12.15	12.88	13.58	15.63
Maximal cliques enrichment in CINs	5.99	1.64	2.22	2.16	1.70
Average radial position of partitions	0.57	0.70	0.60	0.71	0.77
Average size of partitions (number of 200 kb regions)	71.00	32.90	33.28	37.73	59.01
Average number of partitions in each structure	53.86	159.23	91.63	109.79	141.85
Average fraction of inter-chromosome edges in partitions (%)	41.52	25.49	35.26	14.29	9.57

Acknowledgements

This work was supported by the National Institutes of Health (grant U54DK107981 and UM1HG011593 to F.A), and an NSF CAREER grant (1150287 to F.A.). We thank Profs. Andrew Belmont and Jian Ma for useful discussions.

References

1. Chakraborty, A. & Ay, F. The role of 3D genome organization in disease: From compartments to single nucleotides. *Semin Cell Dev Biol* **90**, 104-113 (2019).
2. Bintu, B. et al. Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science* **362**(2018).
3. Wang, S. et al. Spatial organization of chromatin domains and compartments in single chromosomes. *Science* **353**, 598-602 (2016).
4. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-11 (2002).
5. Fang, R. et al. Mapping of long-range chromatin interactions by proximity ligation-assisted ChIP-seq. *Cell Res* **26**, 1345-1348 (2016).
6. Guelen, L. et al. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature* **453**, 948-51 (2008).
7. Hsieh, T.H. et al. Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162**, 108-19 (2015).
8. Li, X. et al. Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. *Nat Protoc* **12**, 899-915 (2017).
9. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-93 (2009).
10. Mumbach, M.R. et al. HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat Methods* **13**, 919-922 (2016).
11. Zheng, M. et al. Multiplex chromatin interactions with single-molecule precision. *Nature* **566**, 558-562 (2019).
12. Schoenfelder, S. et al. Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. *Nat Genet* **42**, 53-61 (2010).
13. Spilianakis, C.G., Lalioti, M.D., Town, T., Lee, G.R. & Flavell, R.A. Interchromosomal associations between alternatively expressed loci. *Nature* **435**, 637-45 (2005).
14. Tolhuis, B., Palstra, R.J., Splinter, E., Grosveld, F. & de Laat, W. Looping and interaction between hypersensitive sites in the active beta-globin locus. *Mol Cell* **10**, 1453-65 (2002).
15. Rao, S.S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).
16. Zhao, R., Bodnar, M.S. & Spector, D.L. Nuclear neighborhoods and gene expression. *Curr Opin Genet Dev* **19**, 172-9 (2009).
17. Chen, Y. et al. Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *J Cell Biol* **217**, 4025-4048 (2018).
18. Vertii, A. et al. Two contrasting classes of nucleolus-associated domains in mouse fibroblast heterochromatin. *Genome Res* **29**, 1235-1249 (2019).
19. Su, J.H., Zheng, P., Kinrot, S.S., Bintu, B. & Zhuang, X. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell* **182**, 1641-1659 e26 (2020).

20. Hua, N. et al. Producing genome structure populations with the dynamic and automated PGS software. *Nat Protoc* **13**, 915-926 (2018).

21. Tjong, H. et al. Population-based 3D genome structure analysis reveals driving forces in spatial genome organization. *Proc Natl Acad Sci U S A* **113**, E1663-72 (2016).

22. Leemans, C. et al. Promoter-Intrinsic and Local Chromatin Features Determine Gene Repression in LADs. *Cell* **177**, 852-864 e14 (2019).

23. Girelli, G. et al. GPSeq reveals the radial organization of chromatin in the cell nucleus. *Nat Biotechnol* **38**, 1184-1193 (2020).

24. Osorio, D., Yu, X., Yu, P., Serpedin, E. & Cai, J.J. Single-cell RNA sequencing of a European and an African lymphoblastoid cell line. *Sci Data* **6**, 112 (2019).

25. Finn, E.H. & Misteli, T. Molecular basis and biological function of variability in spatial genome organization. *Science* **365**(2019).

26. Hildebrand, E.M. & Dekker, J. Mechanisms and Functions of Chromosome Compartmentalization. *Trends Biochem Sci* **45**, 385-396 (2020).

27. Di Pierro, M., Zhang, B., Aiden, E.L., Wolynes, P.G. & Onuchic, J.N. Transferable model for chromosome architecture. *Proc Natl Acad Sci U S A* **113**, 12168-12173 (2016).

28. Di Stefano, M., Paulsen, J., Lien, T.G., Hovig, E. & Micheletti, C. Hi-C-constrained physical models of human chromosomes recover functionally-related properties of genome organization. *Sci Rep* **6**, 35985 (2016).

29. Esposito, A. et al. Polymer physics and machine learning reveal a combinatorial code linking chromatin 3D architecture and 1D epigenetics. *Biorxiv*, 2021 (https://doi.org/10.1101/2021.03.01.433416).

30. Lin, X., Qi, Y., Latham, A.P. & Zhang, B. Multiscale modeling of genome organization with maximum entropy optimization. *J Chem Phys* **155**, 010901 (2021).

31. Paulsen, J. et al. Chrom3D: three-dimensional genome modeling from Hi-C and nuclear lamin-genome contacts. *Genome Biol* **18**, 21 (2017).

32. Qi, Y. et al. Data-Driven Polymer Model for Mechanistic Exploration of Diploid Genome Organization. *Biophys J* **119**, 1905-1916 (2020).

33. Zhang, B. & Wolynes, P.G. Topology, structures, and energy landscapes of human chromosomes. *Proc Natl Acad Sci U S A* **112**, 6062-7 (2015).

34. Li, Q. et al. The three-dimensional genome organization of Drosophila melanogaster through data integration. *Genome Biol* **18**, 145 (2017).

35. Kind, J. et al. Genome-wide maps of nuclear lamina interactions in single human cells. *Cell* **163**, 134-47 (2015).

36. Pope, B.D. et al. Topologically associating domains are stable units of replication-timing regulation. *Nature* **515**, 402-5 (2014).

37. Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F. & Chen, L. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol* **30**, 90-8 (2011).

38. Takizawa, T., Meaburn, K.J. & Misteli, T. The meaning of gene positioning. *Cell* **135**, 9-13 (2008).

39. Bickmore, W.A. The spatial organization of the human genome. *Annu Rev Genomics Hum Genet* **14**, 67-84 (2013).

40. Enright, A.J., Van Dongen, S. & Ouzounis, C.A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**, 1575-84 (2002).

41. Core, L.J. et al. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat Genet* **46**, 1311-20 (2014).

42. Galganski, L., Urbanek, M.O. & Krzyzosiak, W.J. Nuclear speckles: molecular organization, biological function and role in disease. *Nucleic Acids Res* **45**, 10350-10368 (2017).

43. Zhu, L. & Brangwynne, C.P. Nuclear bodies: the emerging biophysics of nucleoplasmic phases. *Curr Opin Cell Biol* **34**, 23-30 (2015).

44. Spector, D.L. & Lamond, A.I. Nuclear speckles. *Cold Spring Harb Perspect Biol* **3**(2011).

45. Brangwynne, C.P., Mitchison, T.J. & Hyman, A.A. Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes. *Proc Natl Acad Sci U S A* **108**, 4334-9 (2011).

46. Hall, L.L., Smith, K.P., Byron, M. & Lawrence, J.B. Molecular anatomy of a speckle. *Anat Rec A Discov Mol Cell Evol Biol* **288**, 664-75 (2006).

47. Shopland, L.S., Johnson, C.V., Byron, M., McNeil, J. & Lawrence, J.B. Clustering of multiple specific genes and gene-rich R-bands around SC-35 domains: evidence for local euchromatic neighborhoods. *J Cell Biol* **162**, 981-90 (2003).

48. Xing, Y., Johnson, C.V., Moen, P.T., Jr., McNeil, J.A. & Lawrence, J. Nonrandom gene organization: structural arrangements of specific pre-mRNA transcription and splicing with SC-35 domains. *J Cell Biol* **131**, 1635-47 (1995).

49. Carter, K.C. et al. A three-dimensional view of precursor messenger RNA metabolism within the mammalian nucleus. *Science* **259**, 1330-5 (1993).

50. Chen, Y. & Belmont, A.S. Genome organization around nuclear speckles. *Curr Opin Genet Dev* **55**, 91-99 (2019).

51. Quinodoz, S.A. et al. Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174**, 744-757 e24 (2018).

52. Kim, J., Venkata, N.C., Hernandez Gonzalez, G.A., Khanna, N. & Belmont, A.S. Gene expression amplification by nuclear speckle association. *J Cell Biol* **219**(2020).

53. Nemeth, A. et al. Initial genomics of the human nucleolus. *PLoS Genet* **6**, e1000889 (2010).

54. Xiong, K. & Ma, J. Revealing Hi-C subcompartments by imputing inter-chromosomal chromatin interactions. *Nat Commun* **10**, 5069 (2019).

55. Ding, F. & Elowitz, M.B. Constitutive splicing and economies of scale in gene expression. *Nat Struct Mol Biol* **26**, 424-432 (2019).

56. Hagberg, A.A., Schult, D.A. & Swart, P.J. “Exploring network structure, dynamics, and function using NetworkX”. in *7th Python in Science Conference (SciPy2008)* (ed. Gäel Varoquaux, T.V., Jarrod Millman) 11-15 (Pasadena, CA USA, 2008).

57. Duarte, M. & Watanabe, R.N. Notes on Scientific Computing for Biomechanics and Motor Control (Version v0.0.2). (http://doi.org/10.5281/zenodo.4599319, 2021).

58. Shin, H. et al. TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res* **44**, e70 (2016).

59. Levy-Leduc, C., Delattre, M., Mary-Huard, T. & Robin, S. Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics* **30**, i386-92 (2014).

60. Crane, E. et al. Condensin-driven remodelling of X chromosome topology during dosage compensation. *Nature* **523**, 240-4 (2015).

61. Serra, F. et al. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput Biol* **13**, e1005665 (2017).

62. Rohatgi, A. WebPlotDigitizer. 4.4 edn (https://automeris.io/WebPlotDigitizer, 2020).

63. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).

64. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).

65. Pettersen, E.F. et al. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-12 (2004).